

US010645380B2

(12) **United States Patent**  
**Xu et al.**

(10) **Patent No.:** **US 10,645,380 B2**  
(45) **Date of Patent:** **May 5, 2020**

(54) **METHOD AND APPARATUS FOR VIDEO CODING**

(71) Applicant: **Tencent America LLC**, Palo Alto, CA (US)

(72) Inventors: **Xiaozhong Xu**, State College, PA (US); **Shan Liu**, San Jose, CA (US); **Meng Xu**, San Jose, CA (US); **Xiang Li**, Los Gatos, CA (US); **Stephan Wenger**, Hillsborough, CA (US)

(73) Assignee: **Tencent America LLC**, Palo Alto, CA (US)

(\*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **16/222,719**

(22) Filed: **Dec. 17, 2018**

(65) **Prior Publication Data**

US 2020/0014917 A1 Jan. 9, 2020

**Related U.S. Application Data**

(60) Provisional application No. 62/695,377, filed on Jul. 9, 2018.

(51) **Int. Cl.**

**H04N 19/105** (2014.01)

**H04N 19/177** (2014.01)

**H04N 19/573** (2014.01)

**H04N 19/139** (2014.01)

**H04N 19/58** (2014.01)

(52) **U.S. Cl.**

CPC ..... **H04N 19/105** (2014.11); **H04N 19/139** (2014.11); **H04N 19/177** (2014.11); **H04N 19/573** (2014.11); **H04N 19/58** (2014.11)

(58) **Field of Classification Search**

None

See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

8,526,495 B2 9/2013 Liu et al.

9,049,452 B2 6/2015 Liu et al.

(Continued)

OTHER PUBLICATIONS

Schwarz et al., "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard," IEEE Trans. on Circuits and Systems for Video Technology, Sep. 2007.\*

(Continued)

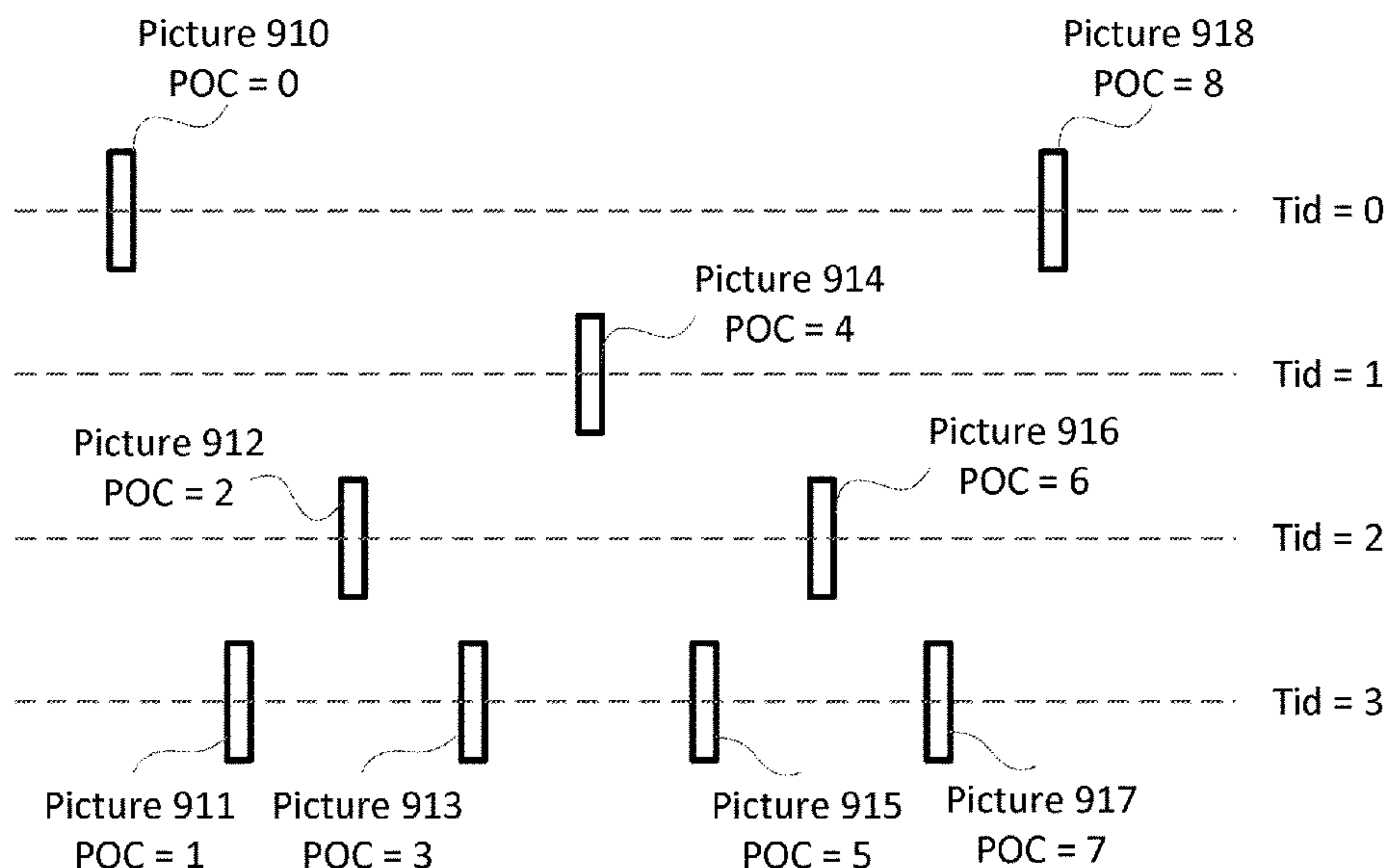
*Primary Examiner* — Michael J Hess

(74) *Attorney, Agent, or Firm* — Oblon, McClelland, Maier & Neustadt, L.L.P.

(57) **ABSTRACT**

Aspects of the disclosure provide method and apparatus for video coding. In some examples, an apparatus includes processing circuitry. The processing circuitry decodes a first portion of video data to obtain first decoded data corresponding to at least two pictures of a plurality of pictures. The processing circuitry identifies one or more pictures of the at least two pictures for decoding a second portion of the video data corresponding to a current picture. In a case that the one or more identified pictures includes two or more identified pictures, the processing circuitry selects a collocated reference picture based on one of (i) the POC numbers of the two or more identified pictures and the current picture, and (ii) a selection index provided in the video data. The processing circuitry also decodes the second portion of the video data using the collocated reference picture.

**19 Claims, 12 Drawing Sheets**



(56)

**References Cited**

U.S. PATENT DOCUMENTS

9,510,012	B2	11/2016	Liu et al.	
9,813,726	B2	11/2017	Liu et al.	
2015/0078446	A1*	3/2015	Jun	H04N 19/593 375/240.12
2015/0156501	A1*	6/2015	Hannuksela	H04N 19/70 375/240.12
2015/0195566	A1*	7/2015	Hinz	H04N 19/105 375/240.13
2015/0264377	A1*	9/2015	He	H04N 19/30 375/240.12
2016/0029029	A1*	1/2016	Lee	H04N 19/187 382/233
2018/0124422	A1*	5/2018	Van Leuven	H04N 19/176
2018/0376160	A1*	12/2018	Zhang	H04N 19/513

OTHER PUBLICATIONS

Shan Liu, et al.; “Nonlinear motion-compensated interpolation for low-bit-rate video”, Proceedings of Spie; SPIEDigitalLibrary.org/conference-proceedings-of-spie; <https://www.spiedigitallibrary.org/conference-proceedings-of-spie> on Apr. 11, 2018 Terms of Use: <https://www.spiedigitallibrary.org/terms-of-use>; International Symposium on Optical Science and Technology, 2000, San Diego, CA, United States; 12 pages.

Shan Liu, et al.; “MCI-embedded Motion Compensated Prediction for Quality Enhancement of Frame Interpolation”; Multimedia Systems and Applications III, Andrew G. Tescher, Bhaskaran Vasudev, V. Michael Bove, Jr., Editors, Proceedings of SPIE vol. 4209 (2001) © 2001 SPIE • 0277-786X/01/\$15.00; <http://proceedings.spiedigitallibrary.org/> on Jul. 3, 2016 Terms of Use: <http://spiedigitallibrary.org/ss/TermsOfUse.aspx> , 11 pages.

Shan Liu, et al.; “Improved Video Coding via Adaptive Selection of Generalized Motion Prediction Modes for B Frames”; Picture Coding Symposium 2001, 4 pages.

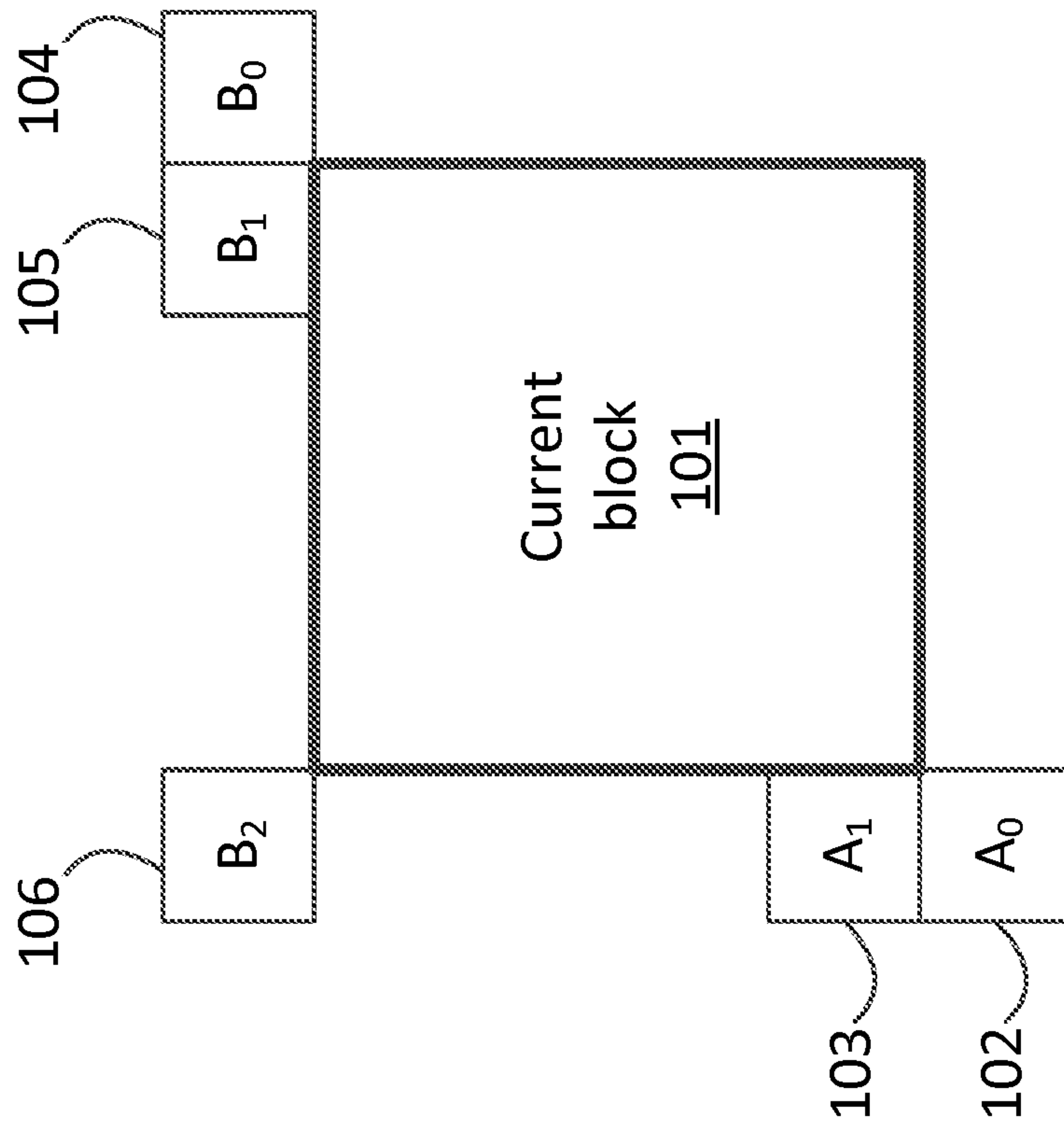
Haitao Yang et al.; “Description of CE4: Inter prediction and motion vector coding”, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, 10<sup>th</sup> Meeting: San Diego, US, Apr. 10-12, 2018, 47 pages.

Xiaozhong Xu et al.; “CE2 Test 3.2: Intra BC merge mode with default candidates”, Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, 20<sup>th</sup> Meeting: Geneva, CH, Feb. 10-18, 2015, 6 pages.

Xiaozhong Xu et al.; “Non-CE2: Intra BC merge mode with default candidates”; Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, 19<sup>th</sup> Meeting: Strasbourg, FR, Oct. 17-24, 2014, 7 pages.

ITU-T; Telecommunication Standardization Sector of ITU; “Series H: Audiovisual and Multimedia Systems”; Infrastructure of audiovisual services—Coding of moving video; High efficiency video coding; H.265 (Dec. 2016), 664 pages.

\* cited by examiner



Related Art

**FIG. 1**

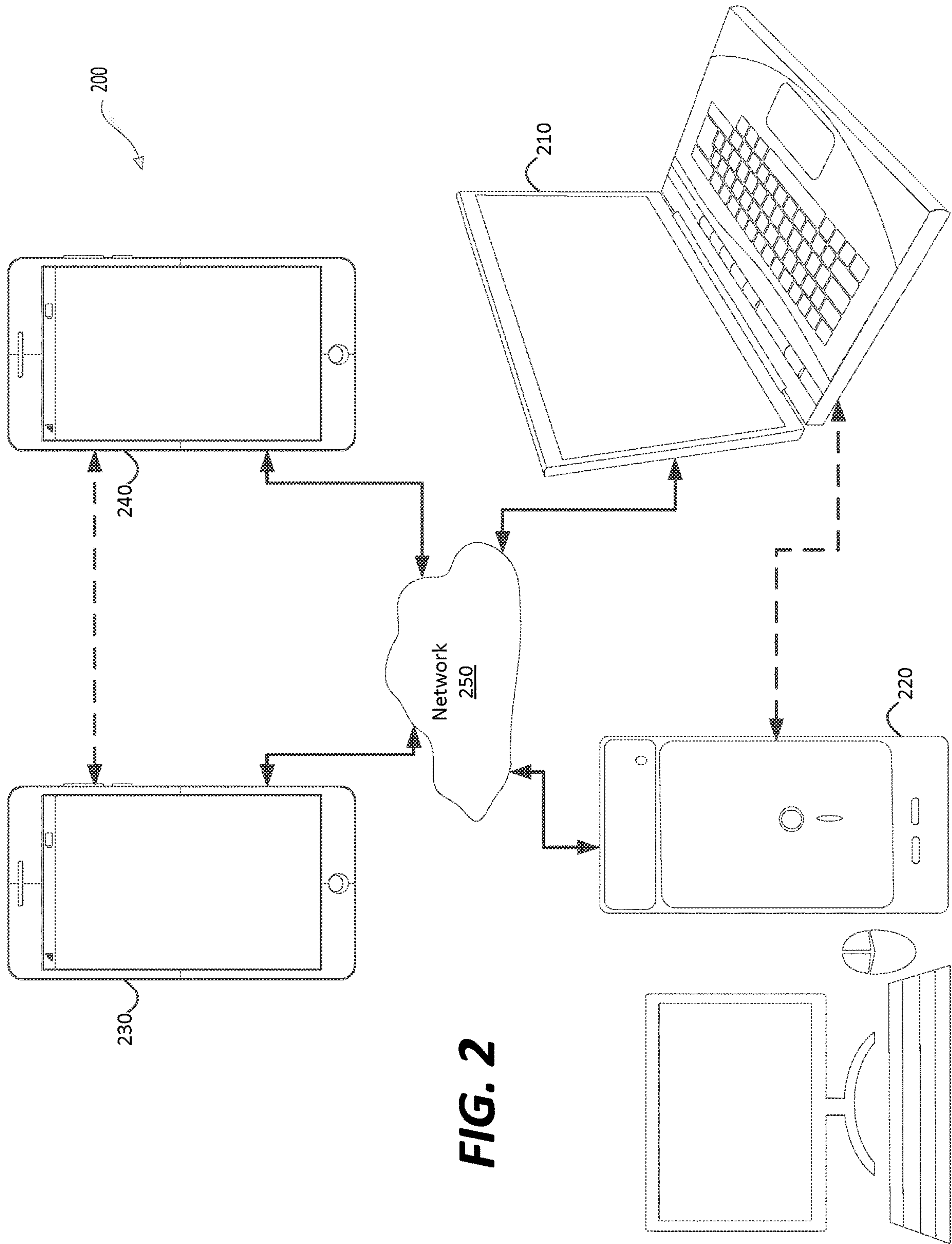
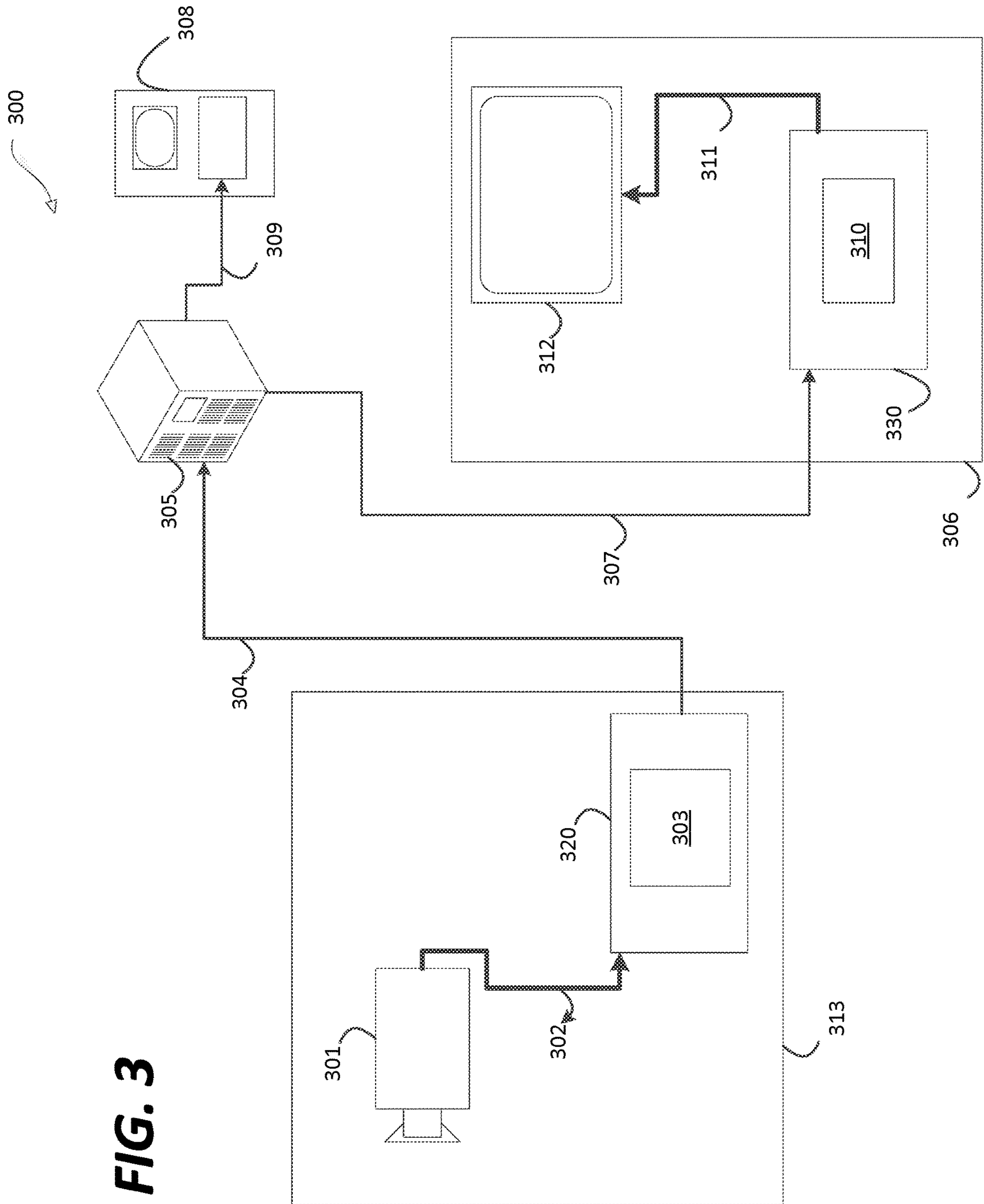


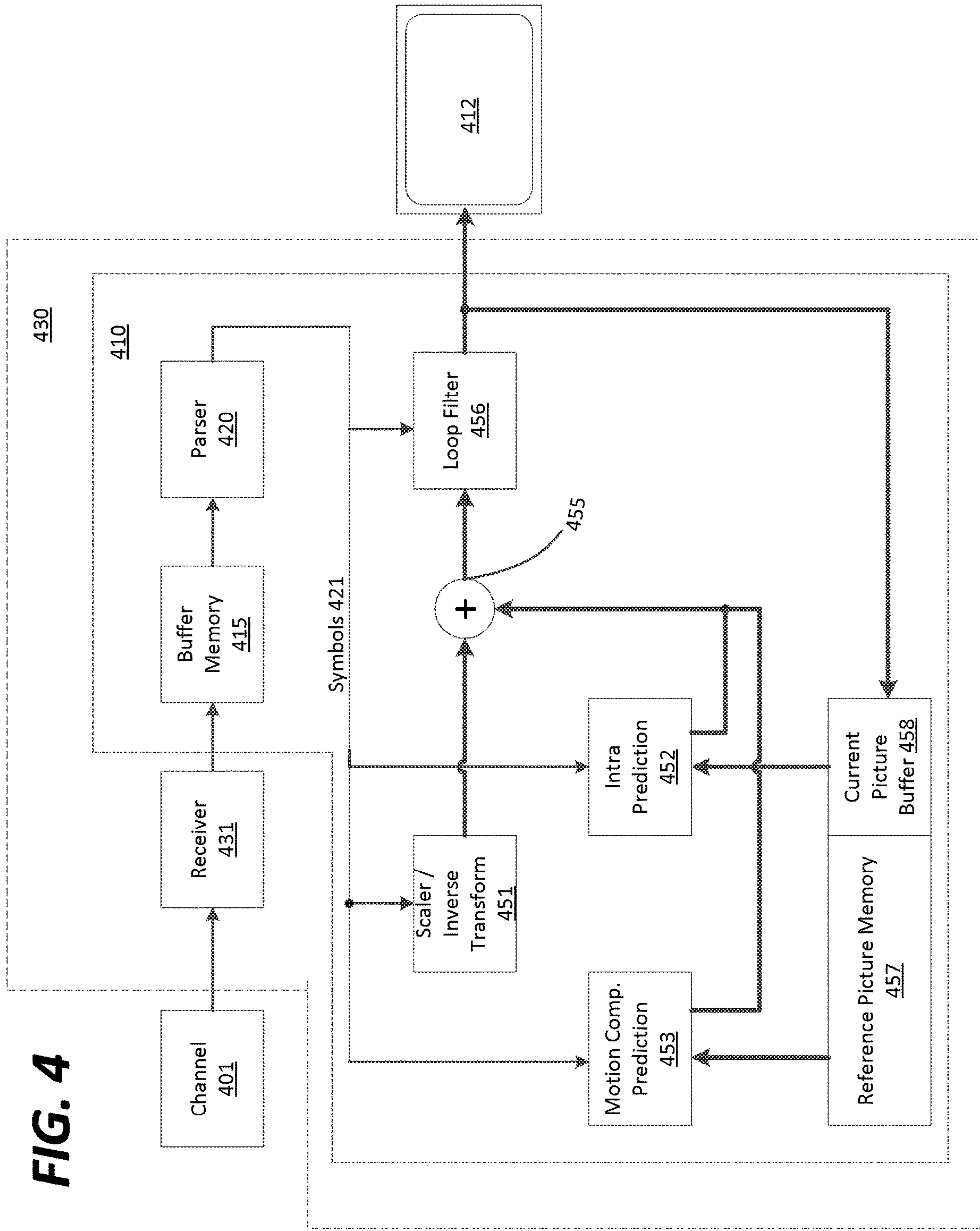
FIG. 2





**FIG. 3**

**FIG. 4**



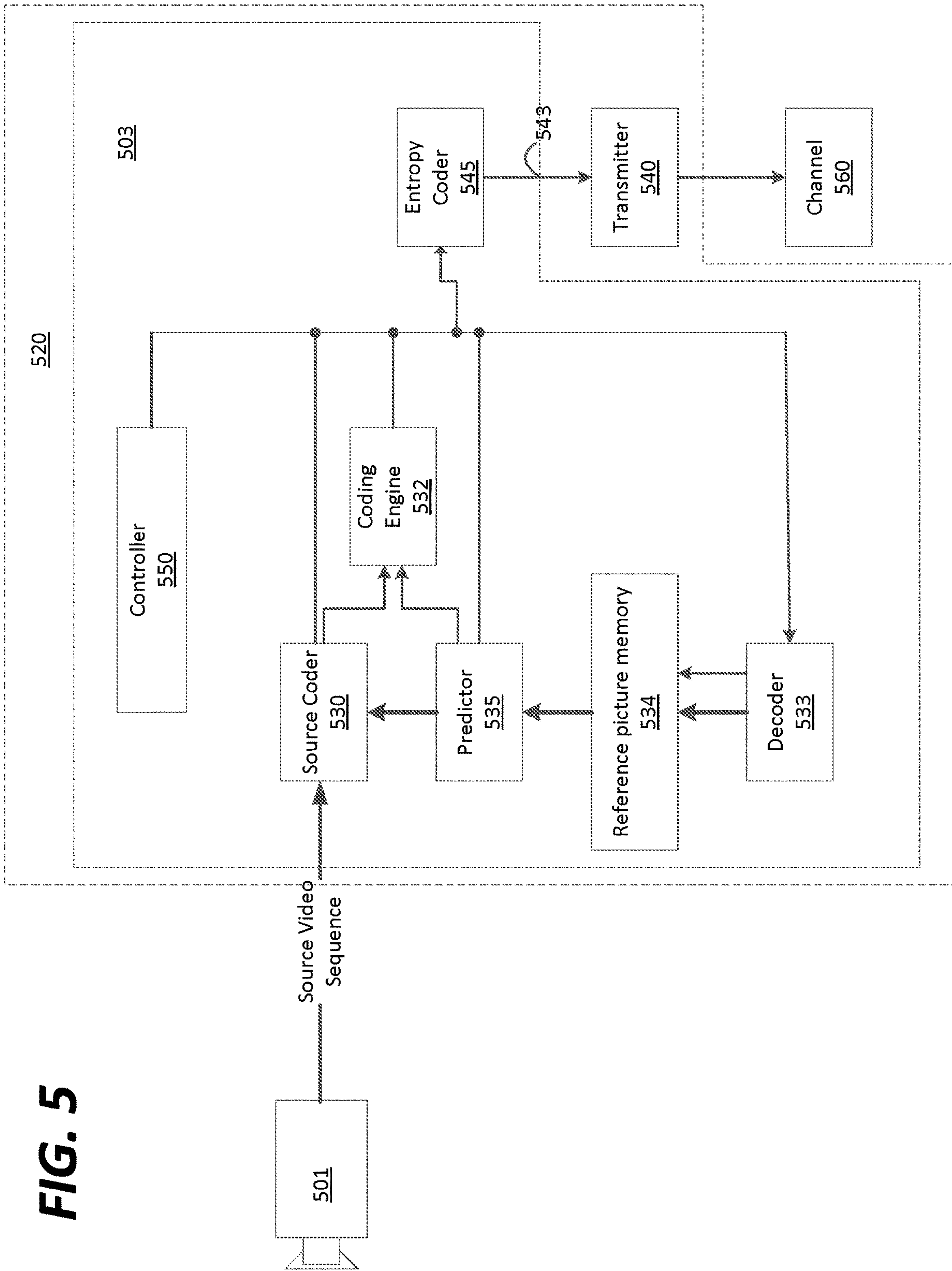
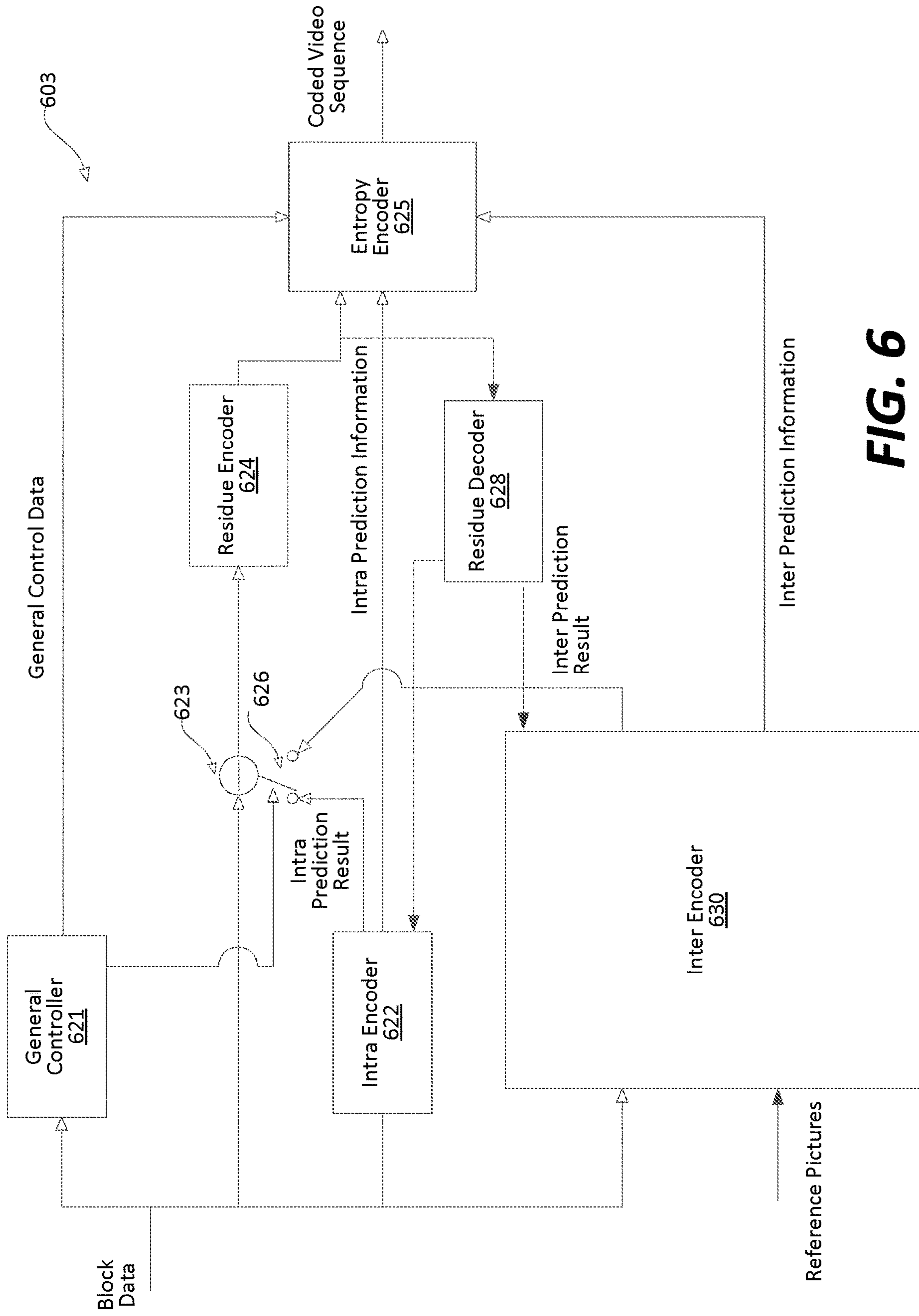
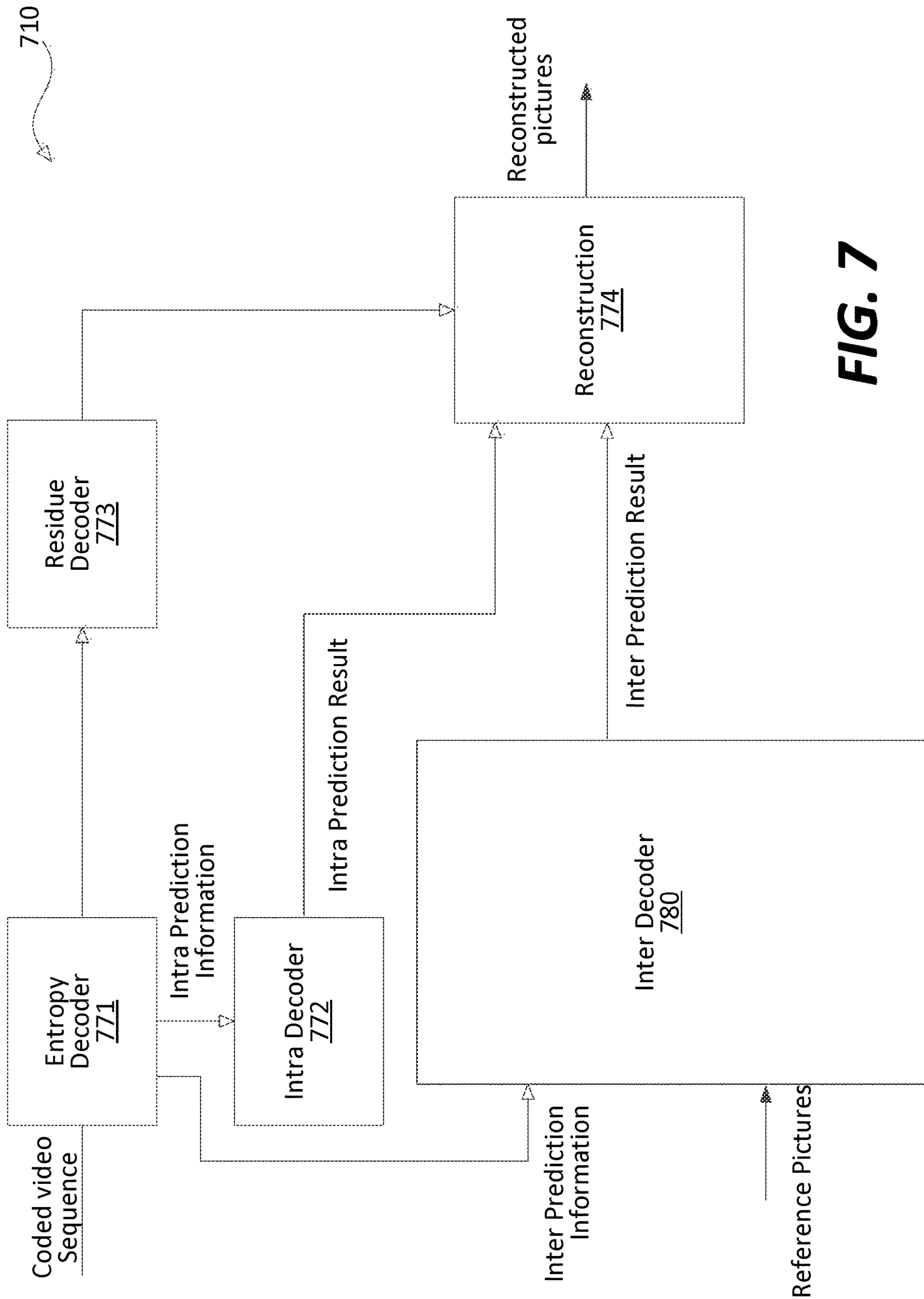


FIG. 5

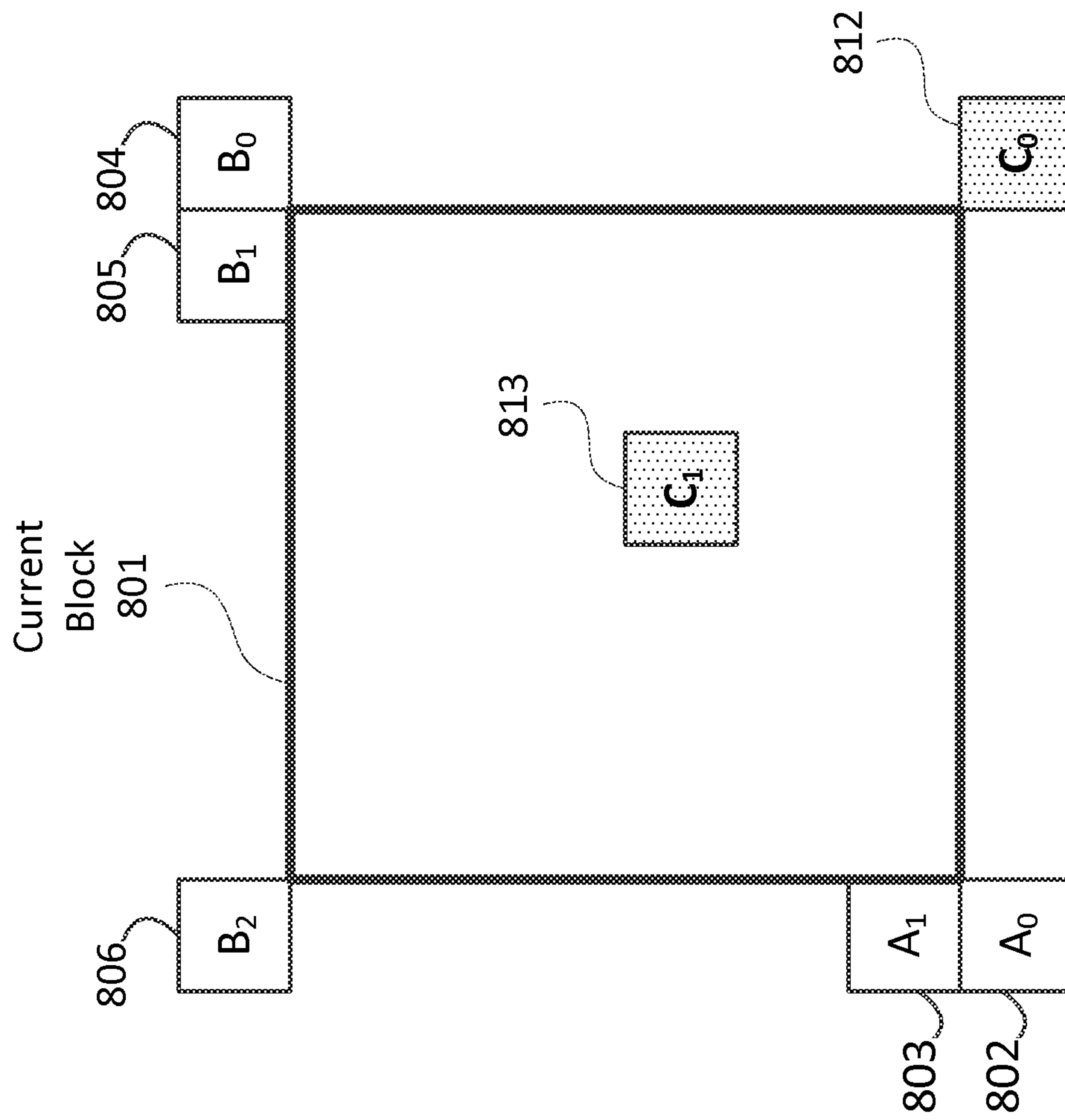


**FIG. 6**

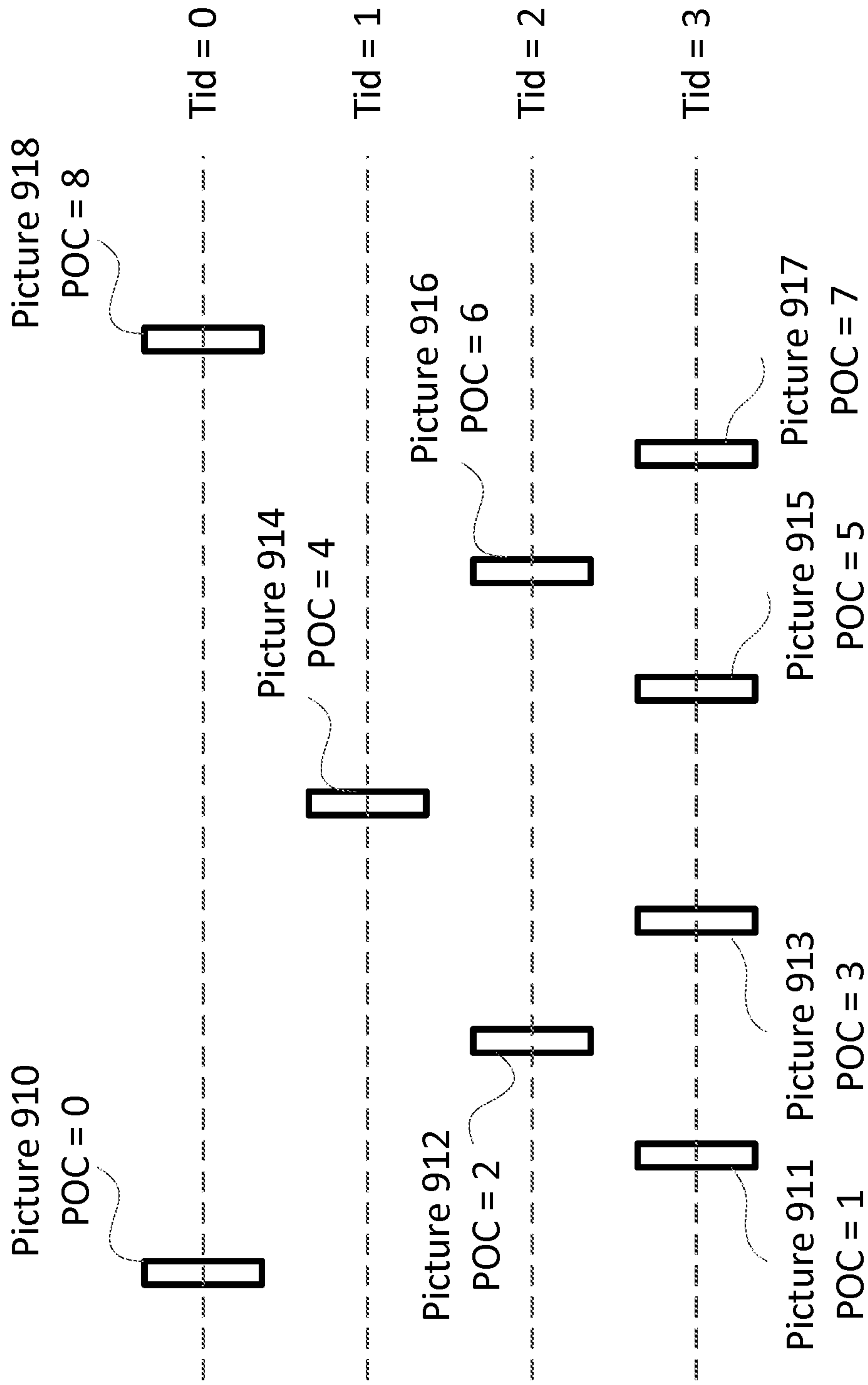




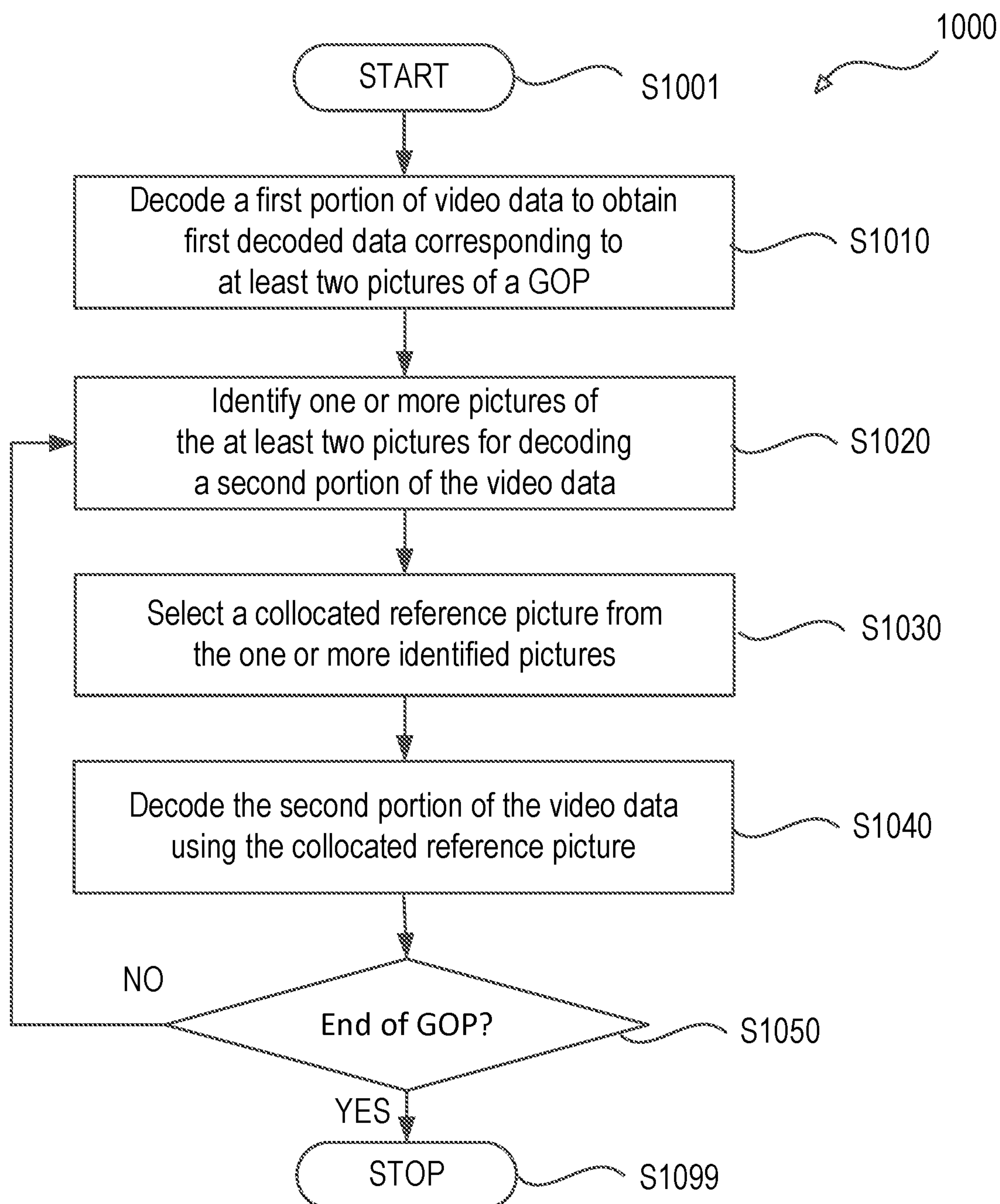
**FIG. 7**



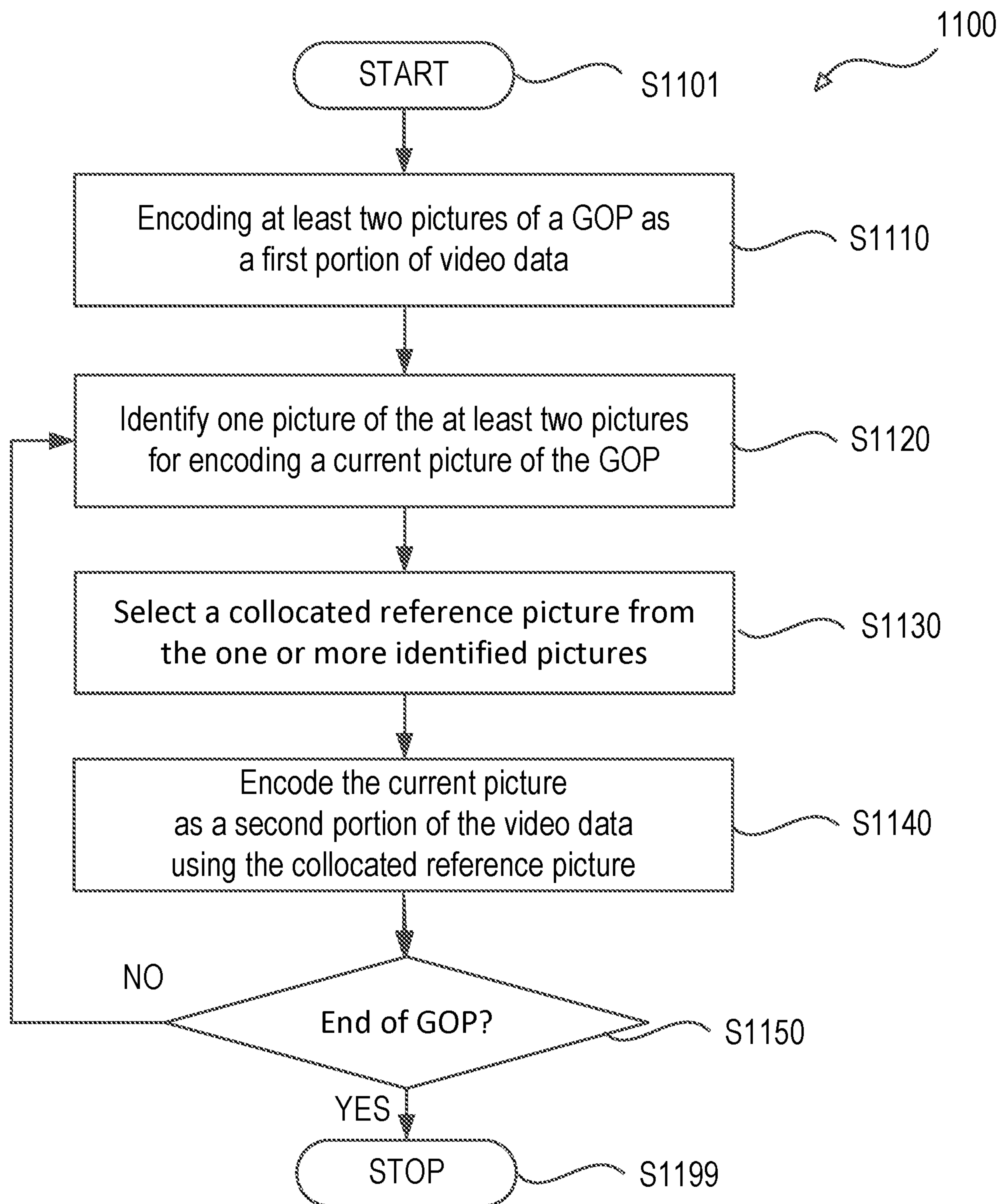
**FIG. 8**



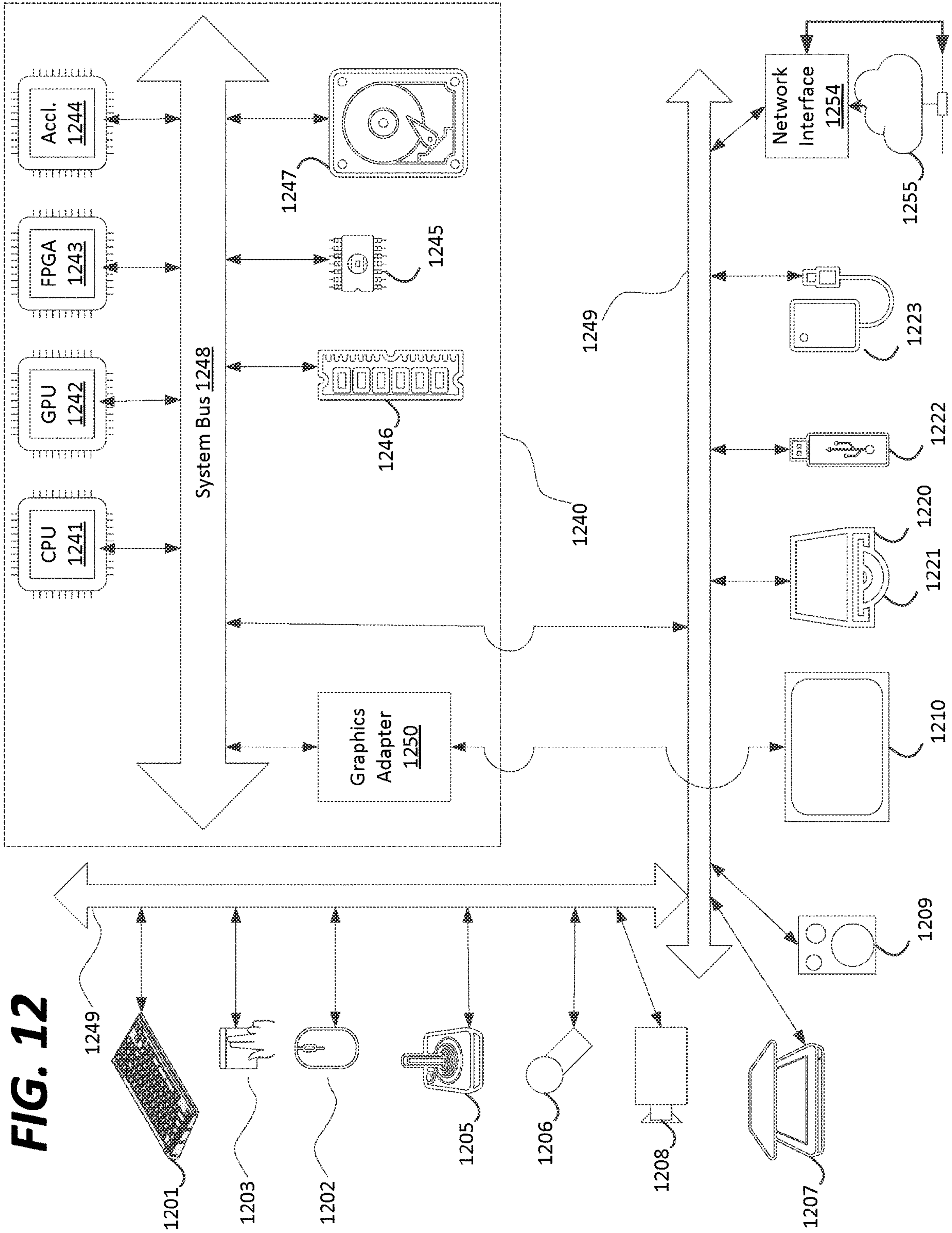
**FIG. 9**

**FIG. 10**





**FIG. 11**





## METHOD AND APPARATUS FOR VIDEO CODING

### INCORPORATION BY REFERENCE

The present disclosure claims the benefit of priority to U.S. Provisional Application No. 62/695,377, "METHODS FOR STORING TEMPORAL MOTION VECTORS IN MOTION COMPENSATION PREDICTION" filed on Jul. 9, 2018, which is incorporated by reference herein in its entirety.

### TECHNICAL FIELD

The present disclosure describes embodiments generally related to video coding.

### BACKGROUND

The background description provided herein is for the purpose of generally presenting the context of the disclosure. Work of the presently named inventors, to the extent the work is described in this background section, as well as aspects of the description that may not otherwise qualify as prior art at the time of filing, are neither expressly nor impliedly admitted as prior art against the present disclosure.

Video coding and decoding can be performed using inter-picture prediction with motion compensation. Uncompressed digital video can include a series of pictures, each picture having a spatial dimension of, for example, 1920×1080 luminance samples and associated chrominance samples. The series of pictures can have a fixed or variable picture rate (informally also known as frame rate), of, for example 60 pictures per second or 60 Hz. Uncompressed video has significant bitrate requirements. For example, 1080p60 4:2:0 video at 8 bit per sample (1920×1080 luminance sample resolution at 60 Hz frame rate) requires close to 1.5 Gbit/s bandwidth. An hour of such video requires more than 600 GBytes of storage space.

One purpose of video coding and decoding can be the reduction of redundancy in the input video signal, through compression. Compression can help reduce the aforementioned bandwidth or storage space requirements, in some cases by two orders of magnitude or more. Both lossless and lossy compression, as well as a combination thereof can be employed. Lossless compression refers to techniques where an exact copy of the original signal can be reconstructed from the compressed original signal. When using lossy compression, the reconstructed signal may not be identical to the original signal, but the distortion between original and reconstructed signals is small enough to make the reconstructed signal useful for the intended application. In the case of video, lossy compression is widely employed. The amount of distortion tolerated depends on the application; for example, users of certain consumer streaming applications may tolerate higher distortion than users of television distribution applications. The compression ratio achievable can reflect that: higher allowable/tolerable distortion can yield higher compression ratios.

Motion compensation can be a lossy compression technique and can relate to techniques where a block of sample data from a previously reconstructed picture or part thereof (reference picture), after being spatially shifted in a direction indicated by a motion vector (MV henceforth), is used for the prediction of a newly reconstructed picture or picture part. In some cases, the reference picture can be the same as

the picture currently under reconstruction. MVs can have two dimensions X and Y, or three dimensions, the third being an indication of the reference picture in use (the latter, indirectly, can be a time dimension).

5 In some video compression techniques, an MV applicable to a certain area of sample data can be predicted from other MVs, for example from those related to another area of sample data spatially adjacent to the area under reconstruction, and preceding that MV in decoding order. Doing so can substantially reduce the amount of data required for coding the MV, thereby removing redundancy and increasing compression. MV prediction can work effectively, for example, because when coding an input video signal derived from a camera (known as natural video) there is a statistical likelihood that areas larger than the area to which a single MV is applicable move in a similar direction and, therefore, can in some cases be predicted using a similar motion vector derived from MVs of neighboring area. That results in the MV found for a given area to be similar or the same as the MV predicted from the surrounding MVs, and that in turn can be represented, after entropy coding, in a smaller number of bits than what would be used if coding the MV directly. In some cases, MV prediction can be an example of lossless compression of a signal (namely: the MVs) derived from the original signal (namely: the sample stream). In other cases, MV prediction itself can be lossy, for example because of rounding errors when calculating a predictor from several surrounding MVs.

Various MV prediction mechanisms are described in H.265/HEVC (ITU-T Rec. H.265, "High Efficiency Video Coding," December 2016). Out of the many MV prediction mechanisms that H.265 offers, described here is a technique henceforth referred to as "spatial merge."

Referring to FIG. 1, a current block (101) comprises samples that have been found by the encoder during the motion search process to be predictable from a previous block of the same size that has been spatially shifted. Instead of coding that MV directly, the MV can be derived from metadata associated with one or more reference pictures, for example from the most recent (in decoding order) reference picture, using the MV associated with either one of five surrounding samples, denoted A0, A1, and B0, B1, B2 (102 through 106, respectively). In H.265, the MV prediction can use predictors from the same reference picture that the neighboring block is using.

### SUMMARY

Aspects of the disclosure provide methods and apparatuses for video coding. In some examples, an apparatus includes processing circuitry that decodes a first portion of video data to obtain first decoded data corresponding to at least two pictures of a plurality of pictures, the video data corresponding to the plurality of pictures that is associated with respective Picture Order Count (POC) numbers indicating a temporal order of the plurality of pictures and respectively in a plurality of temporal layers. The processing circuitry identifies one or more pictures of the at least two pictures for decoding a second portion of the video data corresponding to a current picture. In a case that the one or more identified pictures includes two or more identified pictures, the processing circuitry selects a collocated reference picture from the two or more identified pictures based on one of (i) the POC numbers of the two or more identified pictures and the current picture, and (ii) a selection index provided in the video data, the two or more identified pictures being in different temporal layers. In a case that the



one or more identified pictures includes only one identified picture, the processing circuitry selects the only one identified picture as the collocated reference picture. The processing circuitry also decodes the second portion of the video data using the collocated reference picture to obtain second decoded data corresponding to the current picture.

In some examples, in the case that the one or more identified pictures includes the two or more identified pictures, the processing circuitry selects one of the two or more identified pictures that corresponds to a closest POC number difference with respect to the current picture as the collocated reference picture.

In some examples, in the case that the one or more identified pictures includes the two or more identified pictures, and in a case that two of the two or more identified pictures correspond to a closest POC number difference with respect to the current picture, the processing circuitry performs one of selecting one of the two of the two or more identified pictures in a lowest temporal layer of the identified pictures as the collocated reference picture, and selecting one of the two of the two or more identified pictures in a highest temporal layer of the identified pictures as the collocated reference picture.

In some examples, the plurality of temporal layers includes N temporal layers, and the processing circuitry allocates M memory spaces associated with M of the plurality of temporal layers, respectively, and stores the first decoded data corresponding to the one or more identified pictures in one or more respective memory spaces of the M allocated memory spaces according to the temporal layer of each of the one or more identified pictures, where M is a positive integer ranging from two to N. In some examples, the processing circuitry obtains the positive integer M from the video data. In some examples, the processing circuitry determines the positive integer M based on a picture size of the plurality of pictures. In some embodiments, the positive integer N is greater than two, the positive integer M is less than or equal to the positive integer N, and the M allocated memory spaces are associated with M lower temporal layers among the N plurality of temporal layers.

In some examples, the plurality of temporal layers include N temporal layers, and the processing circuitry allocates M memory spaces, stores the first decoded data corresponding to the one or more identified pictures in one or more respective memory spaces of the M allocated memory spaces, and, in a case that the one or more identified pictures includes less than M pictures and the current picture is in a lowest temporal layer of the temporal layers, stores the second decoded data corresponding to the current picture in a vacant one of the M memory spaces, where M is a positive integer less than or equal to N.

In some examples, in a case that the one or more identified pictures includes a particular picture in a same temporal layer as the current picture, the processing circuitry stores the second decoded data corresponding to the current picture in place of decoded data corresponding to the particular picture. In some examples, in a case that the one or more identified pictures includes a particular picture in a same temporal layer as the current picture, the processing circuitry partially updates decoded data stored in one of the M memory spaces that corresponds to the particular picture using the second decoded data corresponding to the current picture.

Aspects of the disclosure also provide a non-transitory computer-readable medium storing instructions which when executed by a computer for video decoding cause the computer to perform the method for video coding.

#### BRIEF DESCRIPTION OF THE DRAWINGS

Further features, the nature, and various advantages of the disclosed subject matter will be more apparent from the following detailed description and the accompanying drawings in which:

FIG. 1 is a schematic illustration of a current block and its surrounding spatial merge candidates in one example.

FIG. 2 is a schematic illustration of a simplified block diagram of a communication system (200) in accordance with an embodiment.

FIG. 3 is a schematic illustration of a simplified block diagram of a communication system (300) in accordance with an embodiment.

FIG. 4 is a schematic illustration of a simplified block diagram of a decoder in accordance with an embodiment.

FIG. 5 is a schematic illustration of a simplified block diagram of an encoder in accordance with an embodiment.

FIG. 6 shows a block diagram of an encoder in accordance with another embodiment.

FIG. 7 shows a block diagram of a decoder in accordance with another embodiment.

FIG. 8 is a schematic illustration of a current block, its surrounding spatial candidates, and its collocated candidates in one example.

FIG. 9 is a schematic illustration of a plurality of pictures associated with respective Picture Order Count (POC) numbers in a plurality of temporal layers in accordance with an embodiment.

FIG. 10 shows a flow chart outlining a decoding process (1000) according to an embodiment of the disclosure.

FIG. 11 shows a flow chart outlining an encoding process (1100) according to an embodiment of the disclosure.

FIG. 12 is a schematic illustration of a computer system in accordance with an embodiment.

#### DETAILED DESCRIPTION OF EMBODIMENTS

FIG. 2 illustrates a simplified block diagram of a communication system (200) according to an embodiment of the present disclosure. The communication system (200) includes a plurality of terminal devices that can communicate with each other, via, for example, a network (250). For example, the communication system (200) includes a first pair of terminal devices (210) and (220) interconnected via the network (250). In the FIG. 2 example, the first pair of terminal devices (210) and (220) performs unidirectional transmission of data. For example, the terminal device (210) may code video data (e.g., a stream of video pictures that are captured by the terminal device (210)) for transmission to the other terminal device (220) via the network (250). The encoded video data can be transmitted in the form of one or more coded video bitstreams. The terminal device (220) may receive the coded video data from the network (250), decode the coded video data to recover the video pictures and display video pictures according to the recovered video data. Unidirectional data transmission may be common in media serving applications and the like.

In another example, the communication system (200) includes a second pair of terminal devices (230) and (240) that performs bidirectional transmission of coded video data that may occur, for example, during videoconferencing. For bidirectional transmission of data, in an example, each terminal device of the terminal devices (230) and (240) may code video data (e.g., a stream of video pictures that are captured by the terminal device) for transmission to the other terminal device of the terminal devices (230) and (240)



## 5

via the network (250). Each terminal device of the terminal devices (230) and (240) also may receive the coded video data transmitted by the other terminal device of the terminal devices (230) and (240), and may decode the coded video data to recover the video pictures and may display video pictures at an accessible display device according to the recovered video data.

In the FIG. 2 example, the terminal devices (210), (220), (230) and (240) may be illustrated as servers, personal computers, and smart phones, but the principles of the present disclosure may be not so limited. Embodiments of the present disclosure find application with laptop computers, tablet computers, media players, and/or dedicated video conferencing equipment. The network (250) represents any number of networks that convey coded video data among the terminal devices (210), (220), (230), and (240), including for example wireline (wired) and/or wireless communication networks. The communication network (250) may exchange data in circuit-switched and/or packet-switched channels. Representative networks include telecommunications networks, local area networks, wide area networks, and/or the Internet. For the purposes of the present discussion, the architecture and topology of the network (250) may be immaterial to the operation of the present disclosure unless explained herein below.

FIG. 3 illustrates, as an example for an application for the disclosed subject matter, the placement of a video encoder and a video decoder in a streaming environment. The disclosed subject matter can be equally applicable to other video enabled applications, including, for example, video conferencing, digital TV, storing of compressed video on digital media including CD, DVD, memory stick and the like, and so on.

A streaming system may include a capture subsystem (313), that can include a video source (301), for example a digital camera, creating for example a stream of video pictures (302) that are uncompressed. In an example, the stream of video pictures (302) includes samples that are taken by the digital camera. The stream of video pictures (302), depicted as a bold line to emphasize a high data volume when compared to encoded video data (304) (or coded video bitstreams), can be processed by an electronic device (320) that includes a video encoder (303) coupled to the video source (301). The video encoder (303) can include hardware, software, or a combination thereof to enable or implement aspects of the disclosed subject matter as described in more detail below. The encoded video data (304) (or encoded video bitstream (304)), depicted as a thin line to emphasize the lower data volume when compared to the stream of video pictures (302), can be stored on a streaming server (305) for future use. One or more streaming client subsystems, such as client subsystems (306) and (308) in FIG. 3 can access the streaming server (305) to retrieve copies (307) and (309) of the encoded video data (304). A client subsystem (306) can include a video decoder (310), for example, in an electronic device (330). The video decoder (310) decodes the incoming copy (307) of the encoded video data and creates an outgoing stream of video pictures (311) that can be rendered on a display (312) (e.g., display screen) or other rendering device (not depicted). In some streaming systems, the encoded video data (304), (307), and (309) (e.g., video bitstreams) can be encoded according to certain video coding/compression standards. Examples of those standards include ITU-T Recommendation H.265. In an example, a video coding standard under

## 6

development is informally known as Versatile Video Coding (VVC). The disclosed subject matter may be used in the context of VVC.

It is noted that the electronic devices (320) and (330) can include other components (not shown). For example, the electronic device (320) can include a video decoder (not shown) and the electronic device (330) can include a video encoder (not shown) as well.

FIG. 4 shows a block diagram of a video decoder (410) according to an embodiment of the present disclosure. The video decoder (410) can be included in an electronic device (430). The electronic device (430) can include a receiver (431) (e.g., receiving circuitry). The video decoder (410) can be used in the place of the video decoder (310) in the FIG. 3 example.

The receiver (431) may receive one or more coded video sequences to be decoded by the video decoder (410); in the same or another embodiment, one coded video sequence at a time, where the decoding of each coded video sequence is independent from other coded video sequences. The coded video sequence may be received from a channel (401), which may be a hardware/software link to a storage device which stores the encoded video data. The receiver (431) may receive the encoded video data with other data, for example, coded audio data and/or ancillary data streams, that may be forwarded to their respective using entities (not depicted). The receiver (431) may separate the coded video sequence from the other data. To combat network jitter, a buffer memory (415) may be coupled in between the receiver (431) and an entropy decoder/parser (420) (“parser (420)” henceforth). In certain applications, the buffer memory (415) is part of the video decoder (410). In others, it can be outside of the video decoder (410) (not depicted). In still others, there can be a buffer memory (not depicted) outside of the video decoder (410), for example to combat network jitter, and in addition another buffer memory (415) inside the video decoder (410), for example to handle playout timing. When the receiver (431) is receiving data from a store/forward device of sufficient bandwidth and controllability, or from an isosynchronous network, the buffer memory (415) may not be needed, or can be small. For use on best effort packet networks such as the Internet, the buffer memory (415) may be required, can be comparatively large and can be advantageously of adaptive size, and may at least partially be implemented in an operating system or similar elements (not depicted) outside of the video decoder (410).

The video decoder (410) may include the parser (420) to reconstruct symbols (421) from the coded video sequence. Categories of those symbols include information used to manage operation of the video decoder (410), and potentially information to control a rendering device such as a render device (412) (e.g., a display screen) that is not an integral part of the electronic device (430) but can be coupled to the electronic device (430), as was shown in FIG. 4. The control information for the rendering device(s) may be in the form of Supplemental Enhancement Information (SEI messages) or Video Usability Information (VUI) parameter set fragments (not depicted). The parser (420) may parse/entropy-decode the coded video sequence that is received. The coding of the coded video sequence can be in accordance with a video coding technology or standard, and can follow various principles, including variable length coding, Huffman coding, arithmetic coding with or without context sensitivity, and so forth. The parser (420) may extract from the coded video sequence, a set of subgroup parameters for at least one of the subgroups of pixels in the video decoder, based upon at least one parameter corre-



sponding to the group. Subgroups can include Groups of Pictures (GOPs), pictures, tiles, slices, macroblocks, Coding Units (CUs), blocks, Transform Units (TUs), Prediction Units (PUs), and so forth. The parser (420) may also extract from the coded video sequence information such as transform coefficients, quantizer parameter values, motion vectors, and so forth.

The parser (420) may perform an entropy decoding/parsing operation on the video sequence received from the buffer memory (415), so as to create symbols (421).

Reconstruction of the symbols (421) can involve multiple different units depending on the type of the coded video picture or parts thereof (such as: inter and intra picture, inter and intra block), and other factors. Which units are involved, and how, can be controlled by the subgroup control information that was parsed from the coded video sequence by the parser (420). The flow of such subgroup control information between the parser (420) and the multiple units below is not depicted for clarity.

Beyond the functional blocks already mentioned, the video decoder (410) can be conceptually subdivided into a number of functional units as described below. In a practical implementation operating under commercial constraints, many of these units interact closely with each other and can, at least partly, be integrated into each other. However, for the purpose of describing the disclosed subject matter, the conceptual subdivision into the functional units below is appropriate.

A first unit is the scaler/inverse transform unit (451). The scaler/inverse transform unit (451) receives a quantized transform coefficient as well as control information, including which transform to use, block size, quantization factor, quantization scaling matrices, etc. as symbol(s) (421) from the parser (420). The scaler/inverse transform unit (451) can output blocks comprising sample values, that can be input into aggregator (455).

In some cases, the output samples of the scaler/inverse transform (451) can pertain to an intra coded block; that is: a block that is not using predictive information from previously reconstructed pictures, but can use predictive information from previously reconstructed parts of the current picture. Such predictive information can be provided by an intra picture prediction unit (452). In some cases, the intra picture prediction unit (452) generates a block of the same size and shape of the block under reconstruction, using surrounding already reconstructed information fetched from the current picture buffer (458). The current picture buffer (458) buffers, for example, partly reconstructed current picture and/or fully reconstructed current picture. The aggregator (455), in some cases, adds, on a per sample basis, the prediction information the intra prediction unit (452) has generated to the output sample information as provided by the scaler/inverse transform unit (451).

In other cases, the output samples of the scaler/inverse transform unit (451) can pertain to an inter coded, and potentially motion compensated block. In such a case, a motion compensation prediction unit (453) can access reference picture memory (457) to fetch samples used for prediction. After motion compensating the fetched samples in accordance with the symbols (421) pertaining to the block, these samples can be added by the aggregator (455) to the output of the scaler/inverse transform unit (451) (in this case called the residual samples or residual signal) so as to generate output sample information. The addresses within the reference picture memory (457) from where the motion compensation prediction unit (453) fetches prediction samples can be controlled by motion vectors, available to the

motion compensation prediction unit (453) in the form of symbols (421) that can have, for example X, Y, and reference picture components. Motion compensation also can include interpolation of sample values as fetched from the reference picture memory (457) when sub-sample exact motion vectors are in use, motion vector prediction mechanisms, and so forth.

The output samples of the aggregator (455) can be subject to various loop filtering techniques in the loop filter unit (456). Video compression technologies can include in-loop filter technologies that are controlled by parameters included in the coded video sequence (also referred to as coded video bitstream) and made available to the loop filter unit (456) as symbols (421) from the parser (420), but can also be responsive to meta-information obtained during the decoding of previous (in decoding order) parts of the coded picture or coded video sequence, as well as responsive to previously reconstructed and loop-filtered sample values.

The output of the loop filter unit (456) can be a sample stream that can be output to the render device (412) as well as stored in the reference picture memory (457) for use in future inter-picture prediction.

Certain coded pictures, once fully reconstructed, can be used as reference pictures for future prediction. For example, once a coded picture corresponding to a current picture is fully reconstructed and the coded picture has been identified as a reference picture (by, for example, the parser (420)), the current picture buffer (458) can become a part of the reference picture memory (457), and a fresh current picture buffer can be reallocated before commencing the reconstruction of the following coded picture.

The video decoder (410) may perform decoding operations according to a predetermined video compression technology in a standard, such as ITU-T Rec. H.265. The coded video sequence may conform to a syntax specified by the video compression technology or standard being used, in the sense that the coded video sequence adheres to both the syntax of the video compression technology or standard and the profiles as documented in the video compression technology or standard. Specifically, a profile can select certain tools as the only tools available for use under that profile from all the tools available in the video compression technology or standard. Also necessary for compliance can be that the complexity of the coded video sequence is within bounds as defined by the level of the video compression technology or standard. In some cases, levels restrict the maximum picture size, maximum frame rate, maximum reconstruction sample rate (measured in, for example megasamples per second), maximum reference picture size, and so on. Limits set by levels can, in some cases, be further restricted through Hypothetical Reference Decoder (HRD) specifications and metadata for HRD buffer management signaled in the coded video sequence.

In an embodiment, the receiver (431) may receive additional (redundant) data with the encoded video. The additional data may be included as part of the coded video sequence(s). The additional data may be used by the video decoder (410) to properly decode the data and/or to more accurately reconstruct the original video data. Additional data can be in the form of, for example, temporal, spatial, or signal noise ratio (SNR) enhancement layers, redundant slices, redundant pictures, forward error correction codes, and so on.

FIG. 5 shows a block diagram of a video encoder (503) according to an embodiment of the present disclosure. The video encoder (503) is included in an electronic device (520). The electronic device (520) includes a transmitter



(540) (e.g., transmitting circuitry). The video encoder (503) can be used in the place of the video encoder (303) in the FIG. 3 example.

The video encoder (503) may receive video samples from a video source (501) (that is not part of the electronic device (520) in the FIG. 5 example) that may capture video image(s) to be coded by the video encoder (503). In another example, the video source (501) is a part of the electronic device (520).

The video source (501) may provide the source video sequence to be coded by the video encoder (503) in the form of a digital video sample stream that can be of any suitable bit depth (for example: 8 bit, 10 bit, 12 bit, . . . ), any colorspace (for example, BT.601 Y CrCb, RGB, . . . ), and any suitable sampling structure (for example Y CrCb 4:2:0, Y CrCb 4:4:4). In a media serving system, the video source (501) may be a storage device storing previously prepared video. In a videoconferencing system, the video source (501) may be a camera that captures local image information as a video sequence. Video data may be provided as a plurality of individual pictures that impart motion when viewed in sequence. The pictures themselves may be organized as a spatial array of pixels, wherein each pixel can comprise one or more samples depending on the sampling structure, color space, etc. in use. A person skilled in the art can readily understand the relationship between pixels and samples. The description below focuses on samples.

According to an embodiment, the video encoder (503) may code and compress the pictures of the source video sequence into a coded video sequence (543) in real time or under any other time constraints as required by the application. Enforcing appropriate coding speed is one function of a controller (550). In some embodiments, the controller (550) controls other functional units as described below and is functionally coupled to the other functional units. The coupling is not depicted for clarity. Parameters set by the controller (550) can include rate control related parameters (picture skip, quantizer, lambda value of rate-distortion optimization techniques, . . . ), picture size, group of pictures (GOP) layout, maximum motion vector search range, and so forth. The controller (550) can be configured to have other suitable functions that pertain to the video encoder (503) optimized for a certain system design.

In some embodiments, the video encoder (503) is configured to operate in a coding loop. As an oversimplified description, in an example, the coding loop can include a source coder (530) (e.g., responsible for creating symbols, such as a symbol stream, based on an input picture to be coded, and a reference picture(s)), and a (local) decoder (533) embedded in the video encoder (503). The decoder (533) reconstructs the symbols to create the sample data in a similar manner as a (remote) decoder also would create (as any compression between symbols and coded video bit-stream is lossless in the video compression technologies considered in the disclosed subject matter). The reconstructed sample stream (sample data) is input to the reference picture memory (534). As the decoding of a symbol stream leads to bit-exact results independent of decoder location (local or remote), the content in the reference picture memory (534) is also bit exact between the local encoder and remote encoder. In other words, the prediction part of an encoder “sees” as reference picture samples exactly the same sample values as a decoder would “see” when using prediction during decoding. This fundamental principle of reference picture synchronicity (and resulting

drift, if synchronicity cannot be maintained, for example because of channel errors) is used in some related arts as well.

The operation of the “local” decoder (533) can be the same as of a “remote” decoder, such as the video decoder (410), which has already been described in detail above in conjunction with FIG. 4. Briefly referring also to FIG. 4, however, as symbols are available and encoding/decoding of symbols to a coded video sequence by an entropy coder (545) and the parser (420) can be lossless, the entropy decoding parts of the video decoder (410), including the buffer memory (415), and parser (420) may not be fully implemented in the local decoder (533).

An observation that can be made at this point is that any decoder technology except the parsing/entropy decoding that is present in a decoder also necessarily needs to be present, in substantially identical functional form, in a corresponding encoder. For this reason, the disclosed subject matter focuses on decoder operation. The description of encoder technologies can be abbreviated as they are the inverse of the comprehensively described decoder technologies. Only in certain areas a more detail description is required and provided below.

During operation, in some examples, the source coder (530) may perform motion compensated predictive coding, which codes an input picture predictively with reference to one or more previously-coded picture from the video sequence that were designated as “reference pictures”. In this manner, the coding engine (532) codes differences between pixel blocks of an input picture and pixel blocks of reference picture(s) that may be selected as prediction reference(s) to the input picture.

The local video decoder (533) may decode coded video data of pictures that may be designated as reference pictures, based on symbols created by the source coder (530). Operations of the coding engine (532) may advantageously be lossy processes. When the coded video data may be decoded at a video decoder (not shown in FIG. 5), the reconstructed video sequence typically may be a replica of the source video sequence with some errors. The local video decoder (533) replicates decoding processes that may be performed by the video decoder on reference pictures and may cause reconstructed reference pictures to be stored in the reference picture cache (534). In this manner, the video encoder (503) may store copies of reconstructed reference pictures locally that have common content as the reconstructed reference pictures that will be obtained by a far-end video decoder (absent transmission errors).

The predictor (535) may perform prediction searches for the coding engine (532). That is, for a new picture to be coded, the predictor (535) may search the reference picture memory (534) for sample data (as candidate reference pixel blocks) or certain metadata such as reference picture motion vectors, block shapes, and so on, that may serve as an appropriate prediction reference for the new pictures. The predictor (535) may operate on a sample block-by-pixel block basis to find appropriate prediction references. In some cases, as determined by search results obtained by the predictor (535), an input picture may have prediction references drawn from multiple reference pictures stored in the reference picture memory (534).

The controller (550) may manage coding operations of the source coder (530), including, for example, setting of parameters and subgroup parameters used for encoding the video data.

Output of all aforementioned functional units may be subjected to entropy coding in the entropy coder (545). The



entropy coder (545) translates the symbols as generated by the various functional units into a coded video sequence, by lossless compressing the symbols according to technologies such as Huffman coding, variable length coding, arithmetic coding, and so forth.

The transmitter (540) may buffer the coded video sequence(s) as created by the entropy coder (545) to prepare for transmission via a communication channel (560), which may be a hardware/software link to a storage device which would store the encoded video data. The transmitter (540) may merge coded video data from the video coder (503) with other data to be transmitted, for example, coded audio data and/or ancillary data streams (sources not shown).

The controller (550) may manage operation of the video encoder (503). During coding, the controller (550) may assign to each coded picture a certain coded picture type, which may affect the coding techniques that may be applied to the respective picture. For example, pictures often may be assigned as one of the following picture types:

An Intra Picture (I picture) may be one that may be coded and decoded without using any other picture in the sequence as a source of prediction. Some video codecs allow for different types of intra pictures, including, for example Independent Decoder Refresh (“IDR”) Pictures. A person skilled in the art is aware of those variants of I pictures and their respective applications and features.

A predictive picture (P picture) may be one that may be coded and decoded using intra prediction or inter prediction using at most one motion vector and reference index to predict the sample values of each block.

A bi-directionally predictive picture (B Picture) may be one that may be coded and decoded using intra prediction or inter prediction using at most two motion vectors and reference indices to predict the sample values of each block. Similarly, multiple-predictive pictures can use more than two reference pictures and associated metadata for the reconstruction of a single block.

Source pictures commonly may be subdivided spatially into a plurality of sample blocks (for example, blocks of 4×4, 8×8, 4×8, or 16×16 samples each) and coded on a block-by-block basis. Blocks may be coded predictively with reference to other (already coded) blocks as determined by the coding assignment applied to the blocks’ respective pictures. For example, blocks of I pictures may be coded non-predictively or they may be coded predictively with reference to already coded blocks of the same picture (spatial prediction or intra prediction). Pixel blocks of P pictures may be coded predictively, via spatial prediction or via temporal prediction with reference to one previously coded reference picture. Blocks of B pictures may be coded predictively, via spatial prediction or via temporal prediction with reference to one or two previously coded reference pictures.

The video encoder (503) may perform coding operations according to a predetermined video coding technology or standard, such as ITU-T Rec. H.265. In its operation, the video encoder (503) may perform various compression operations, including predictive coding operations that exploit temporal and spatial redundancies in the input video sequence. The coded video data, therefore, may conform to a syntax specified by the video coding technology or standard being used.

In an embodiment, the transmitter (540) may transmit additional data with the encoded video. The source coder (530) may include such data as part of the coded video sequence. Additional data may comprise temporal/spatial/SNR enhancement layers, other forms of redundant data

such as redundant pictures and slices, SEI messages, VUI parameter set fragments, and so on.

A video may be captured as a plurality of source pictures (video pictures) in a temporal sequence. Intra-picture prediction (often abbreviated to intra prediction) makes use of spatial correlation in a given picture, and inter-picture prediction makes uses of the (temporal or other) correlation between the pictures. In an example, a specific picture under encoding/decoding, which is referred to as a current picture, is partitioned into blocks. When a block in the current picture is similar to a reference block in a previously coded and still buffered reference picture in the video, the block in the current picture can be coded by a vector that is referred to as a motion vector. The motion vector points to the reference block in the reference picture, and can have a third dimension identifying the reference picture, in case multiple reference pictures are in use.

In some embodiments, a bi-prediction technique can be used in the inter-picture prediction. According to the bi-prediction technique, two reference pictures, such as a first reference picture and a second reference picture that are both prior in decoding order to the current picture in the video (but may be in the past and future, respectively, in display order) are used. A block in the current picture can be coded by a first motion vector that points to a first reference block in the first reference picture, and a second motion vector that points to a second reference block in the second reference picture. The block can be predicted by a combination of the first reference block and the second reference block.

Further, a merge mode technique can be used in the inter-picture prediction to improve coding efficiency.

According to some embodiments of the disclosure, predictions, such as inter-picture predictions and intra-picture predictions are performed in the unit of blocks. For example, according to the HEVC standard, a picture in a sequence of video pictures is partitioned into coding tree units (CTU) for compression, the CTUs in a picture have the same size, such as 64×64 pixels, 32×32 pixels, or 16×16 pixels. In general, a CTU includes three coding tree blocks (CTBs), which are one luma CTB and two chroma CTBs. Each CTU can be recursively quadtree split into one or multiple coding units (CUs). For example, a CTU of 64×64 pixels can be split into one CU of 64×64 pixels, or 4 CUs of 32×32 pixels, or 16 CUs of 16×16 pixels. In an example, each CU is analyzed to determine a prediction type for the CU, such as an inter prediction type or an intra prediction type. The CU is split into one or more prediction units (PUs) depending on the temporal and/or spatial predictability. Generally, each PU includes a luma prediction block (PB), and two chroma PBs. In an embodiment, a prediction operation in coding (encoding/decoding) is performed in the unit of a prediction block. Using a luma prediction block as an example of a prediction block, the prediction block includes a matrix of values (e.g., luma values) for pixels, such as 8×8 pixels, 16×16 pixels, 8×16 pixels, 16×8 pixels, and the like.

FIG. 6 shows a diagram of a video encoder (603) according to another embodiment of the disclosure. The video encoder (603) is configured to receive a processing block (e.g., a prediction block) of sample values within a current video picture in a sequence of video pictures, and encode the processing block into a coded picture that is part of a coded video sequence. In an example, the video encoder (603) is used in the place of the video encoder (303) in the FIG. 3 example.

In an HEVC example, the video encoder (603) receives a matrix of sample values for a processing block, such as a prediction block of 8×8 samples, and the like. The video



encoder (603) determines whether the processing block is best coded using intra mode, inter mode, or bi-prediction mode using, for example, rate-distortion optimization. When the processing block is to be coded in intra mode, the video encoder (603) may use an intra prediction technique to encode the processing block into the coded picture; and when the processing block is to be coded in inter mode or bi-prediction mode, the video encoder (603) may use an inter prediction or bi-prediction technique, respectively, to encode the processing block into the coded picture. In certain video coding technologies, merge mode can be an inter picture prediction submode where the motion vector is derived from one or more motion vector predictors without the benefit of a coded motion vector component outside the predictors. In certain other video coding technologies, a motion vector component applicable to the subject block may be present. In an example, the video encoder (603) includes other components, such as a mode decision module (not shown) to determine the mode of the processing blocks.

In the FIG. 6 example, the video encoder (603) includes the inter encoder (630), an intra encoder (622), a residue calculator (623), a switch (626), a residue encoder (624), a general controller (621), and an entropy encoder (625) coupled together as shown in FIG. 6.

The inter encoder (630) is configured to receive the samples of the current block (e.g., a processing block), compare the block to one or more reference blocks in reference pictures (e.g., blocks in previous pictures and later pictures), generate inter prediction information (e.g., description of redundant information according to inter encoding technique, motion vectors, merge mode information), and calculate inter prediction results (e.g., predicted block) based on the inter prediction information using any suitable technique. In some examples, the reference pictures are decoded reference pictures that are decoded based on the encoded video information.

The intra encoder (622) is configured to receive the samples of the current block (e.g., a processing block), in some cases compare the block to blocks already coded in the same picture, generate quantized coefficients after transform, and in some cases also intra prediction information (e.g., an intra prediction direction information according to one or more intra encoding techniques). In an example, the intra encoder (622) also calculates intra prediction results (e.g., predicted block) based on the intra prediction information and reference blocks in the same picture.

The general controller (621) is configured to determine general control data and control other components of the video encoder (603) based on the general control data. In an example, the general controller (621) determines the mode of the block, and provides a control signal to the switch (626) based on the mode. For example, when the mode is the intra mode, the general controller (621) controls the switch (626) to select the intra mode result for use by the residue calculator (623), and controls the entropy encoder (625) to select the intra prediction information and include the intra prediction information in the bitstream; and when the mode is the inter mode, the general controller (621) controls the switch (626) to select the inter prediction result for use by the residue calculator (623), and controls the entropy encoder (625) to select the inter prediction information and include the inter prediction information in the bitstream.

The residue calculator (623) is configured to calculate a difference (residue data) between the received block and prediction results selected from the intra encoder (622) or the inter encoder (630). The residue encoder (624) is configured to operate based on the residue data to encode the

residue data to generate the transform coefficients. In an example, the residue encoder (624) is configured to convert the residue data from a spatial domain to a frequency domain, and generate the transform coefficients. The transform coefficients are then subject to quantization processing to obtain quantized transform coefficients. In various embodiments, the video encoder (603) also includes a residue decoder (628). The residue decoder (628) is configured to perform inverse-transform, and generate the decoded residue data. The decoded residue data can be suitably used by the intra encoder (622) and the inter encoder (630). For example, the inter encoder (630) can generate decoded blocks based on the decoded residue data and inter prediction information, and the intra encoder (622) can generate decoded blocks based on the decoded residue data and the intra prediction information. The decoded blocks are suitably processed to generate decoded pictures and the decoded pictures can be buffered in a memory circuit (not shown) and used as reference pictures in some examples.

The entropy encoder (625) is configured to format the bitstream to include the encoded block. The entropy encoder (625) is configured to include various information according to a suitable standard, such as the HEVC standard. In an example, the entropy encoder (625) is configured to include the general control data, the selected prediction information (e.g., intra prediction information or inter prediction information), the residue information, and other suitable information in the bitstream. Note that, according to the disclosed subject matter, when coding a block in the merge submode of either inter mode or bi-prediction mode, there is no residue information.

FIG. 7 shows a diagram of a video decoder (710) according to another embodiment of the disclosure. The video decoder (710) is configured to receive coded pictures that are part of a coded video sequence, and decode the coded pictures to generate reconstructed pictures. In an example, the video decoder (710) is used in the place of the video decoder (310) in the FIG. 3 example.

In the FIG. 7 example, the video decoder (710) includes an entropy decoder (771), an inter decoder (780), a residue decoder (773), a reconstruction module (774), and an intra decoder (772) coupled together as shown in FIG. 7.

The entropy decoder (771) can be configured to reconstruct, from the coded picture, certain symbols that represent the syntax elements of which the coded picture is made up. Such symbols can include, for example, the mode in which a block is coded (such as, for example, intra mode, inter mode, bi-predicted mode, the latter two in merge submode or another submode), prediction information (such as, for example, intra prediction information or inter prediction information) that can identify certain sample or metadata that is used for prediction by the intra decoder (772) or the inter decoder (780), respectively, residual information in the form of, for example, quantized transform coefficients, and the like. In an example, when the prediction mode is inter or bi-predicted mode, the inter prediction information is provided to the inter decoder (780); and when the prediction type is the intra prediction type, the intra prediction information is provided to the intra decoder (772). The residual information can be subject to inverse quantization and is provided to the residue decoder (773).

The inter decoder (780) is configured to receive the inter prediction information, and generate inter prediction results based on the inter prediction information.

The intra decoder (772) is configured to receive the intra prediction information, and generate prediction results based on the intra prediction information.



The residue decoder (773) is configured to perform inverse quantization to extract de-quantized transform coefficients, and process the de-quantized transform coefficients to convert the residual from the frequency domain to the spatial domain. The residue decoder (773) may also require certain control information (to include the Quantizer Parameter (QP)), and that information may be provided by the entropy decoder (771) (data path not depicted as this may be low volume control information only).

The reconstruction module (774) is configured to combine, in the spatial domain, the residual as output by the residue decoder (773) and the prediction results (as output by the inter or intra prediction modules as the case may be) to form a reconstructed block, that may be part of the reconstructed picture, which in turn may be part of the reconstructed video. It is noted that other suitable operations, such as a deblocking operation and the like, can be performed to improve the visual quality.

It is noted that the video encoders (303), (503), and (603), and the video decoders (310), (410), and (710) can be implemented using any suitable technique. In an embodiment, the video encoders (303), (503), and (603), and the video decoders (310), (410), and (710) can be implemented using one or more integrated circuits. In another embodiment, the video encoders (303), (503), and (603), and the video decoders (310), (410), and (710) can be implemented using one or more processors that execute software instructions.

FIG. 8 is a schematic illustration of a current block (801), its surrounding spatial candidates, and its collocated candidates in one example.

Referring to FIG. 8, the motion information of a current block (801) can be derived based on the motion information of the surrounding blocks, denoted as surrounding spatial candidates A0, A1, and B0, B1, B2 (for blocks 802 through 806), respectively. Moreover, in some applications, the motion information of the current block (801) can be derived based on the motion information of a predetermined collocated candidate (also referred to as a temporal candidate). For example, the current block (801) may have a collocated block in a specified reference picture. If block (812) of the specified reference picture that corresponds to a position outside the current block and adjacent to a lower-right corner of the current block (801) is coded using the inter-picture prediction, the block (812) is used as the collocated block and the motion information thereof is used as the collocated candidate (C0). However, if the block (812) is not coded using the inter-picture prediction, block (813) that corresponds to a position at the lower-right side of, and adjacent to, a center of the block (801) is used as the collocated block and the motion information thereof is used as the collocated candidate (C1). In some examples, at least one pruning operation can be performed to ensure duplicated candidates are not included in a candidate list more than once.

In some video coding standards, after using spatial and temporal motion information of neighboring blocks to predict the motion information of the current block, the prediction residue is further coded. Such a method is referred to in HEVC as Advanced Motion Vector Prediction (AMVP) mode.

In some examples, a two-candidate motion vector predictor list can be formed. The first candidate predictor is from the first available motion vector from the left edge, in the order of spatial A0, A1 candidates. The second candidate predictor is from the first available motion vector from the top edge, in the order of spatial B0, B1 and B2 candidates. If no valid motion vector can be found from the checked

locations for either the left edge or the top edge, no candidate will be filled in the list. If two candidates are determined to be available and are the same, only one will be kept in the list. If the list is not full (i.e., with two different candidates), the temporal collocated motion vector (after scaling), e.g., collocated candidate (C0), can be used as another candidate. If candidate (C0) is not available, collocated candidate (C1) can be used instead. In some examples, if there are still not enough motion vector predictor candidates after checking the collocated candidates (C0) and (C1), a zero motion vector can be added to the list.

FIG. 9 is a schematic illustration of a plurality of pictures associated with respective Picture Order Count (POC) numbers in a plurality of temporal layers in accordance with an embodiment.

In order to support temporal scalability, in some video coding configurations, the pictures in a sequence are encoded/decoded in a reordered manner. For example, FIG. 9 shows a plurality of pictures (910 through 918) in a group of pictures (GOP). The pictures are associated with respective POC numbers (POC=1 through 8) indicating a temporal order of the pictures and respectively in a plurality of temporal layers. Each layer of the temporal layers is identified in FIG. 9 by a respective temporal ID (Tid). In some examples, when decoding, pictures in lower temporal layers (with smaller Tids) will be decoded prior to decoding picture in higher temporal layers (with larger Tids). The reference pictures may be assigned to facilitate such decoding ordering. For example, a picture in a higher temporal layer can use pictures in a lower temporal layer or pictures of the same temporal layer as reference pictures. However, a picture in a lower temporal layer cannot use a picture in a higher temporal layer as a reference picture.

In FIG. 9, picture (910) is associated with POC=0 and is in a temporal layer with Tid=0; picture (911) is associated with POC=1 and is in a temporal layer with Tid=3; picture (912) is associated with POC=2 and is in a temporal layer with Tid=2; picture (913) is associated with POC=3 and is in the temporal layer with Tid=3; picture (914) is associated with POC=4 and is in the temporal layer with Tid=1; picture (915) is associated with POC=5 and is in the temporal layer with Tid=3; picture (916) is associated with POC=6 and is in the temporal layer with Tid=2; picture (917) is associated with POC=7 and is in the temporal layer with Tid=3; and picture (918) is associated with POC=8 and is in the temporal layer with Tid=0. In some examples, pictures (910-918) can have a decoding order listed using the POC numbers thereof: 0, 8, 4, 2, 1, 3, 6, 5, 7.

In some applications, such as deriving motion information using the merge mode or the AMVP mode, the temporal motion information, which includes motion vectors and other information (such as prediction direction, reference index, etc.) for one or more previously decoded pictures (or also referred to as reference pictures), would be required to be stored such that when coding a block in a future picture, such temporal vectors can be used as a motion vector predictor.

In some coding methods, a slice header flag is used to specify which one of the reference pictures is used for the current slice for deriving the temporal motion vector predictors. This picture is referred as the "collocated reference picture." In some coding methods, because any previously decoded picture can be assigned as the "collocated reference picture" for decoding a current picture, all decoded pictures within the slice and associated motion information are all stored in a decoded picture buffer.



In a video coding system, for each reference picture stored in the decoded picture buffer, a picture-size memory space is allocated to store the image samples of the reference picture. At the same time, another memory space (also called a motion information memory space) associated with this picture-size memory space is also allocated to store the motion vectors and other motion information in this picture. To store K reference pictures, K picture-size memory spaces and K motion information memory spaces are required. Reducing the number of reference pictures that need to be stored can reduce the required memory spaces and improve memory channel bandwidth efficiency.

In some embodiments, the reference pictures that can be assigned as a collocated reference picture for a current picture is limited to at most one decoded picture for each temporal layer. In some examples, the number of total motion information memory spaces needed is decided by the number of temporal layers in the coding structure. In some examples, when the pictures in the highest temporal layer cannot be used as a reference picture, the amount of total motion vector (and motion information) memory space needed is decided by the number of temporal layers in the coding structure minus 1. In some examples, for pictures of each temporal layer, the motion information memory space can store the motion vector and other motion information of the most recently encoded/decoded picture in this temporal layer.

Using the GOP shown in FIG. 9 as an example, the pictures are arranged into a four-temporal-layer structure. For each layer, a motion information memory space is allocated and assigned to store temporal motion vectors and motion information for the most recently encoded/decoded picture of this layer. As shown in Table I, the available collocated reference picture(s) for each picture (in decoding order) varies as the decoding process progresses.

TABLE I

Available collocated reference picture according to temporal layers									
	Decoding Order								
	0	1	2	3	4	5	6	7	8
“Current” POC	0	8	4	2	1	3	6	5	7
Tid	0	0	1	2	3	3	2	3	3
Available Collocated POC	—	0	8	4, 8	2, 4, 8	2, 4, 8	2, 4, 6, 8	4, 6, 8	4, 6, 8

As shown in Table I, after picture (910) (POC=0, Decoding Order=0) is decoded, the motion information memory space associated with temporal layer (Tid=0) will store the motion information associated with this picture (910) for decoding the next picture (918). After picture (918) (POC=8, Decoding Order=1) is decoded, the motion information memory space associated with temporal layer (Tid=0) will store the motion information associated with this picture (918) in place of the motion information of the picture (910) for decoding the next picture (914). After picture (914) (POC=4, Decoding Order=2) is decoded, the motion information memory space associated with (Tid=1) will store the motion information associated with this picture (914). Therefore, for decoding picture (912), motion information of pictures (918) and (914) may be stored and selectable.

After picture (912) (POC=2, Decoding Order=3) is decoded, the motion information memory space associated with (Tid=2) will store the motion information associated

with this picture (912). After picture (916) (POC=6, Decoding Order=6) is decoded, the motion information memory space information memory space (Tid=2) will be updated with the motion information associated with this picture (916). If the pictures in the temporal layer (Tid=3), such as pictures (911, 913, 915, and 917) can be used as reference pictures, a motion information memory space for the temporal layer (Tid=3) will also be allocated and assigned and updated accordingly.

In Table I, for decoding some pictures, such as picture (916) (POC=6), more than one reference picture may qualify to be used as the collocated reference picture, such as pictures (912), (914), and (918) (POC=2, 4, 8). The choice of collocated reference picture may be signaled using control information included in the encoded video data, or derived based on information such as a POC difference and/or temporal layer relationship.

In one example, the collocated reference picture is selected based on a selection index provided in the video data.

In some examples, the one of the available reference pictures that corresponds to a closest POC number difference with respect to the current picture is selected as the collocated reference picture.

In some examples, when in a case that the available reference pictures include two pictures with a same closest POC number difference with respect to the current picture, the one of the two pictures in a lower temporal layer of the two pictures can be selected as the collocated reference picture (as shown in Table II), or the one of the two pictures in a higher temporal layer of the two pictures can be selected as the collocated reference picture (as shown in Table III).

TABLE II

	Collocated picture selection according to POC difference and higher temporal layer								
	Decoding Order								
	0	1	2	3	4	5	6	7	8
“Current” POC	0	8	4	2	1	3	6	5	7
Tid	0	0	1	2	3	3	2	3	3
Collocated picture POC	—	0	8	4	2	2	4	4	6

TABLE III

	Collocated picture selection according to POC difference and lower temporal layer								
	Decoding Order								
	0	1	2	3	4	5	6	7	8
“Current” POC	0	8	4	2	1	3	6	5	7
Tid	0	0	1	2	3	3	2	3	3
Collocated picture POC	—	0	8	4	2	4	8	6	8

In some embodiments, the reference pictures that can be assigned as a collocated reference picture for a current picture is limited to at most a maximum number M of decoded pictures, which may be in M lower temporal layers among the plurality of temporal layers. In some examples,



the number of total motion information memory spaces needed is decided by the number of temporal layers in the coding structure and a buffer size threshold. Also, in some examples, the number of total motion information memory spaces needed is capped by a threshold which specifies the maximum allowed buffer size. For example, if the number of temporal layers is  $N$  and the buffer size threshold determines that a maximum number of motion information memory spaces for different temporal layers is  $M$ , where  $M$  is a positive integer equal to or less than  $N$ , then only the pictures in the lower  $M$  temporal layers will be associated with a respective motion information memory space. In some examples, for pictures of each temporal layer, the motion information memory space can store the motion vector and other motion information of the most recently encoded/decoded picture in this temporal layer.

In one embodiment, the maximum number of motion information memory spaces (e.g., the positive integer  $M$ ) can be determined based on a picture size of the plurality of pictures. In some examples, compared with a set of higher resolution pictures, the maximum number of motion information memory spaces for a set of lower resolution pictures can be set at a greater number.

In another embodiment, the maximum number of motion information memory spaces (e.g., the positive integer  $M$ ) can be determined based on a maximum supported picture resolution. In some examples, when the decoder is set to process pictures at a lower resolution, compared with processing the pictures at a higher, default, resolution, the maximum number of motion information memory spaces can be set at a greater number.

Using the GOP shown in FIG. 9 as an example, the pictures are arranged into a four-temporal-layer structure. For each layer, a motion information memory space is allocated and assigned to store temporal motion vectors and motion information for the most recently encoded/decoded picture of this layer. Also, in some examples, if the maximum allowed motion information memory spaces is set to two, only pictures in the temporal layers ( $Tid=0$ ) and ( $Tid=1$ ) will be used as available collocated reference pictures. As shown in Table IV, the available collocated reference picture(s) for each picture (in decoding order) varies as the decoding process progresses.

TABLE IV

Available collocated reference picture according to $M = 2$ lower temporal layers									
Decoding Order									
	0	1	2	3	4	5	6	7	8
“Current” POC	0	8	4	2	1	3	6	5	7
Tid	0	0	1	2	3	3	2	3	3
Available Collocated POC	—	0	8	4, 8	4, 8	4, 8	4, 8	4, 8	4, 8

As shown in Table IV, after picture (918) (POC=8, Decoding Order=1) is decoded, the motion information memory space associated with temporal layer ( $Tid=0$ ) will store the motion information associated with this picture (918) for decoding the next picture (914). After picture (914) (POC=4, Decoding Order=2) is decoded, the motion information memory space associated with ( $Tid=1$ ) will store the motion information associated with this picture (914). In this example, for all subsequently decoded pictures with  $Tid$

greater than one, the motion information memory spaces will not be updated and the motion information corresponding to pictures (918) and (914) will remain as the available collocated reference pictures.

In Table IV, for decoding some pictures, pictures (912), (911), (913), (916), (915), and (917) (POC=2, 1, 3, 6, 5, 7) may have more than one available collocated reference picture, such as pictures (914) and (918) (POC=4, 8). The choice of collocated reference picture may be signaled using control information included in the encoded video data, or derived based on information such as a POC difference and/or temporal layer relationship.

In one example, the collocated reference picture is selected based on a selection index provided in the video data.

In some examples, the one of the available reference pictures that corresponds to a closest POC number difference with respect to the current picture is selected as the collocated reference picture.

In some examples, when the available reference pictures include two pictures with a same closest POC number difference with respect to the current picture, the one of the two pictures in a lower temporal layer of the two pictures can be selected as the collocated reference picture (as shown in Table V), or the one of the two pictures in a higher temporal layer of the two pictures can be selected as the collocated reference picture (as shown in Table VI). See, for example, the columns for decoding picture (916) (POC=6) in Table V and Table VI.

TABLE V

Collocated picture selection according to POC difference and higher temporal layer									
Decoding Order									
	0	1	2	3	4	5	6	7	8
“Current” POC	0	8	4	2	1	3	6	5	7
Tid	0	0	1	2	3	3	2	3	3
Collocated picture POC	—	0	8	4	4	4	4	4	8

TABLE VI

Collocated picture selection according to POC difference and lower temporal layer									
Decoding Order									
	0	1	2	3	4	5	6	7	8
“Current” POC	0	8	4	2	1	3	6	5	7
Tid	0	0	1	2	3	3	2	3	3
Collocated picture POC	—	0	8	4	4	4	8	4	8

In some other embodiments, when the selection of the collocated picture is specified at a slice level for corresponding slices, there is no need to signal the selection of collocated picture at a picture level.

In some embodiments, the reference pictures that can be assigned as a collocated reference picture for a current picture is limited to at most a maximum number  $M$  of decoded pictures, which may be in  $M$  lower temporal layers



among the plurality of temporal layers, where the motion information of two pictures in a lowest temporal layer of the temporal layers may be concurrently stored in a case that not all M allocated motion information memory spaces are occupied.

In some examples, the number of temporal layers is N and the buffer size threshold determines that a maximum number of motion information memory spaces for different temporal layers is M, where M is a positive integer equal to or less than N. When the number of decoded pictures having decoded data stored in the motion information memory spaces is less than M (i.e., the allocated motion information memory spaces are not all occupied by decoded data of respective pictures), the motion information of a decoded picture in the lowest temporal layer of the temporal layers is allowed to be stored in a vacant motion information memory space, even when the motion information of another picture in the lowest temporal layer has already been stored in one of the allocated motion information memory spaces. However, when the number of decoded pictures having decoded data stored in the motion information memory spaces is M (i.e., the allocated motion information memory spaces are all occupied by decoded data of respective pictures), the stored pictures and the currently decoded picture are in different temporal layers, and the stored pictures include two pictures in the lowest temporal layer, the decoded data of the currently decoded picture can still be stored but in one of the two motion information memory spaces that is occupied by one of the two pictures in the lowest temporal layer that has a smallest POC number. Afterwards, the motion information stored in the M motion information memory spaces will correspond to M pictures in M different temporal layers.

Using the GOP shown in FIG. 9 as an example, the pictures are arranged into a four-temporal-layer structure, where M is set to three. As shown in Table VII, the available collocated reference picture(s) for each picture (in decoding order) varies as the decoding process progresses.

TABLE VII

Available collocated reference picture according to temporal layers									
	Decoding Order								
	0	1	2	3	4	5	6	7	8
"Current" POC	0	8	4	2	1	3	6	5	7
Tid	0	0	1	2	3	3	2	3	3
Available Collocated POC	—	0	0,	0,	2,	2,	2,	4,	4,
			8	4,	4,	4,	4,	6,	6,
				8	8	8	8	8	8

As shown in Table VII, after picture (910) (POC=0, Decoding Order=0) is decoded, the motion information associated with this picture (910) is stored in one of M memory spaces for decoding the next picture (918). After picture (918) (POC=8, Decoding Order=1) is decoded, because at least one vacant memory space is available, the motion information associated with this picture (918) is stored in another one of the M memory spaces for decoding the next picture (914) without clearing the stored motion information associated with picture (910). At this stage, motion information of two pictures in the lowest temporal layer is stored in the motion information memory spaces. After picture (914) (POC=4, Decoding Order=2) is decoded, because one vacant memory space is still available, the motion information associated with this picture (914) is stored in yet another one of the M memory spaces for

decoding the next picture (912) without clearing the stored motion information associated with pictures (910) and (918). Therefore, for decoding picture (912), motion information of pictures (910), (918), and (914) may be stored and selectable.

After picture (912) (POC=2, Decoding Order=3) is decoded, the motion information associated with this picture (912) will be stored to update the motion information associated with picture (910), which has the lowest POC among the stored pictures in the lowest temporal layer.

In some embodiments, various approaches are applicable to update stored motion information associated with a particular temporal layer.

In some examples, the stored motion information can be replaced in a picture-wise manner. For example, in a case that the stored motion information of a particular picture and a currently decoded picture are in a same temporal layer, the motion information of the currently decoded picture is stored in place of the motion information of particular picture. Using Table VII as an example, after picture (912) (POC=2, Decoding Order=3) is decoded, the entirety of stored motion information associated with picture (910) (POC=0) is deleted and replaced with the motion information associated with picture (912) (POC=2).

In some other examples, the stored motion information can be replaced in a block-wise manner. For example, in a case that the stored motion information of a particular picture and a currently decoded picture are in a same temporal layer, the stored motion information of particular picture can be partially updated by the motion information of the currently decoded picture. Using Table VII as an example, after picture (912) (POC=2, Decoding Order=3) is decoded, the stored motion information associated with picture (910) (POC=0) is updated by the motion information associated with picture (912) (POC=2). Specifically, in some examples, for each block, when a block of the picture (912) (POC=2) contains motion information, the motion information for a collocated block in the picture (910) (POC=0) is replaced by the motion information of this block of the picture (912) (POC=2). After all blocks in the picture (910) (POC=0) have been processed, the motion information memory space will be indicated as associated with the picture (912) (POC=2).

The choice of collocated reference picture may be signaled using control information included in the encoded video data, or derived based on information such as a POC difference and/or temporal layer relationship. In one example, the collocated reference picture is selected based on a selection index provided in the video data. In some examples, the one of the available reference pictures that corresponds to a closest POC number difference with respect to the current picture is selected as the collocated reference picture.

In some examples, when the available reference pictures include two pictures with a same closest POC number difference with respect to the current picture, the one of the two pictures in a lower temporal layer of the two pictures can be selected as the collocated reference picture (as shown in Table VIII), or the one of the two pictures in a higher temporal layer of the two pictures can be selected as the collocated reference picture (as shown in Table IX).



TABLE VIII

Collocated picture selection according to POC difference and higher temporal layer									
	Decoding Order								
	0	1	2	3	4	5	6	7	8
“Current” POC	0	8	4	2	1	3	6	5	7
Tid	0	0	1	2	3	3	2	3	3
Collocated picture POC	—	0	0	0	2	2	4	4	6

TABLE IX

Collocated picture selection according to POC difference and lower temporal layer									
	Decoding Order								
	0	1	2	3	4	5	6	7	8
“Current” POC	0	8	4	2	1	3	6	5	7
Tid	0	0	1	2	3	3	2	3	3
Collocated picture POC	—	0	0	4	2	4	8	6	8

FIG. 10 shows a flow chart outlining a decoding process (1000) according to an embodiment of the disclosure. The process (1000) can be used in the reconstruction of a plurality of pictures, such as a GOP, having blocks coded in inter mode. In some embodiments, one or more operations may be performed before or after process (1000), and some of the operations illustrated in FIG. 10 may be reordered or omitted.

In various embodiments, the process (1000) is executed by processing circuitry, such as the processing circuitry in the terminal devices (210), (220), (230), and (240), the processing circuitry that performs functions of the video decoder (310), (410), or (710), and the like. In some embodiments, the process (1000) is implemented in software instructions, thus when the processing circuitry executes the software instructions, the processing circuitry performs the process (1000). The process starts at (S1001) and proceeds to (S1010).

At (S1010), a first portion of video data is decoded to obtain first decoded data corresponding to at least two pictures of a plurality of pictures (e.g., GOP). In some examples, the video data corresponding to the plurality of pictures is associated with respective Picture Order Count (POC) numbers indicating a temporal order of the plurality of pictures and respectively in a plurality of temporal layers, as illustrated with reference to FIG. 9. For example, as shown in Table I, after decoding at least pictures (910) and (918) (POC=0.8), a portion of video data is decoded to obtain decoded data, including motion information, that corresponds to the decoded pictures. In some examples, the first portion of video data can be decoded using the system or decoders illustrated with reference to FIGS. 3, 4, and 7.

At (S1020), one or more pictures of the at least two pictures are identified for decoding a second portion of the video data corresponding to a current picture. For example, based on the rules illustrated with reference to Table I, Table IV, and Table VII, at different stages of decoding a GOP, one

or more pictures that can be used as available collocated reference pictures are identified.

In some embodiments, the plurality of temporal layers includes N temporal layers. Before performing (S1020), M memory spaces associated with M of the plurality of temporal layers can be allocated, respectively, where M is a positive integer less than or equal to N, or in some examples ranging from two to N. In some examples, the first decoded data corresponding to the one or more identified pictures can be stored in one or more respective memory spaces of the M allocated memory spaces according to the temporal layer of each of the one or more identified pictures.

In some examples, the positive integer M can be obtained from the video data. In some examples, the positive integer M can be determined based on a picture size of the plurality of pictures.

In some embodiments, the positive integer N is greater than two, the positive integer M is less than or equal to the positive integer N, and the M allocated memory spaces are associated with M lower temporal layers among the N plurality of temporal layers.

At (S1030), a collocated reference picture is selected from the one or more identified pictures. In some examples, in a case that the one or more identified pictures includes two or more identified pictures, a collocated reference picture can be selected from the two or more identified pictures based on one of (i) the POC numbers of the two or more identified pictures and the current picture, and (ii) a selection index provided in the video data. The two or more identified pictures are in different temporal layers. In some other examples, in a case that the one or more identified pictures includes only one identified picture, the only one identified picture can be selected as the collocated reference picture.

In some examples, one of the two or more identified pictures with a closest POC number difference with respect to the current picture can be selected as the collocated reference picture. In some examples, in a case that there are two of the two or more identified pictures with a closest POC number difference with respect to the current picture, the collocated reference picture can be selected from these two pictures corresponding to the same POC number difference by either (i) selecting one of the two of the two or more identified pictures in a lowest temporal layer of the identified pictures as the collocated reference picture, or (ii) selecting one of the two of the two or more identified pictures in a highest temporal layer of the identified pictures as the collocated reference picture, as similarly illustrated with reference to Table II, Table III, Table V, Table VI, and Table VIII, and Table IX.

At (S1040), the second portion of the video data is decoded using the selected collocated reference picture to obtain second decoded data corresponding to the current picture. In some examples, the second portion of video data can be decoded using the system or decoders illustrated in FIGS. 3, 4, and 7.

At (S1050), in a case that all pictures in the GOP are decoded, the process proceeds to (S1099); and in a case that not all pictures in the GOP are decoded, the process proceeds to (S1020). With the newly decoded data from (S1040) corresponding to the current picture, at (S1020), whether the current picture from (S1040) is included in the newly identified pictures for decoding a next picture is determined.

In some examples, in a case that the one or more previously identified pictures includes less than M pictures and the current picture is in a lowest temporal layer of the temporal layers, the second decoded data corresponding to



the current picture is identified as an available collocated reference picture and is stored in a vacant one of the M memory spaces.

In some examples, in a case that the one or more previously identified pictures includes M pictures, the one or more previously identified pictures and the current picture are in different temporal layers, and the one or more previously identified pictures includes two pictures in the lowest temporal layer, the second decoded data corresponding to the current picture is stored in one of the M memory spaces that stores decoded data corresponding to one of the two pictures in the lowest temporal layer that has a smallest POC number.

In some examples, in a case that the one or more previously identified pictures includes a particular picture in a same temporal layer as the current picture, the second decoded data corresponding to the current picture is stored in place of decoded data corresponding to the particular picture. In some alternative examples, in a case that the one or more previously identified pictures includes a particular picture in a same temporal layer as the current picture, decoded data stored in one of the M memory spaces that corresponds to the particular picture is partially updated using the second decoded data corresponding to the current picture.

At (S1099), the process for decoding the POC terminates.

FIG. 11 shows a flow chart outlining an encoding process (1100) according to an embodiment of the disclosure. The process (1100) can be used in encoding a plurality of pictures, such as a GOP, having blocks coded in inter mode. In some embodiments, one or more operations may be performed before or after process (1100), and some of the operations illustrated in FIG. 10 may be reordered or omitted.

In various embodiments, the process (1100) is executed by processing circuitry, such as the processing circuitry in the terminal devices (210), (220), (230), and (240), the processing circuitry that performs functions of the video encoder (303), (503), or (603), and the like. In some embodiments, the process (1100) is implemented in software instructions, thus when the processing circuitry executes the software instructions, the processing circuitry performs the process (1100). The process starts at (S1101) and proceeds to (S1110).

At (S1110), at least two pictures of a GOP are encoded as a first portion of video data. In some examples, the GOP includes a plurality of pictures that is associated with respective Picture Order Count (POC) numbers indicating a temporal order of the plurality of pictures and respectively in a plurality of temporal layers, as illustrated with reference to FIG. 9. In some examples, the pictures can be encoded using the system or encoders illustrated in FIGS. 3, 5, and 6.

At (S1120), one or more pictures of the at least two pictures are identified for encoding a current picture. For example, based on the rules illustrated with reference to Table I, Table IV, and Table VII, and similarly illustrated with reference to (S1020), at different stages of encoding a GOP, one or more pictures that can be used as available collocated reference pictures are identified.

At (S1130), a collocated reference picture is selected from the one or more identified pictures. In some examples, in a case that the one or more identified pictures includes two or more identified pictures, a collocated reference picture can be selected from the two or more identified pictures based on one of (i) the POC numbers of the two or more identified pictures and the current picture, and (ii) evaluating coding

complexity and/or efficiency of using different collocated reference picture. In the scenario (i), the collocated reference picture can be selected in a manner similar to various examples illustrated with reference to (S1030) and Table II, Table III, Table V, Table VI, and Table VIII, and Table IX. In the scenario (ii), the selected collocated reference picture based on the coding complexity and/or efficiency is to be provided as a selection index included in the encoded video data.

At (S1140), the current picture can be encoded as a second portion of the video data using the collocated reference picture. In some examples, the second portion of video data can be encoded using the system or encoders illustrated in FIGS. 3, 5, and 6.

At (S1150), in a case that all pictures in the GOP are encoded, the process proceeds to (S1199); and in a case that not all pictures in the GOP are encoded, the process proceeds to (S1120). After encoding the current picture, at (S1120), whether the current picture from (S1140) is included in the newly identified pictures for decoding a next picture is determined in a manner similar to the examples as illustrated with reference to (S1150) and (S1120).

At (S1199), the process for encoding the POC terminates.

The techniques described above, can be implemented as computer software using computer-readable instructions and physically stored in one or more computer-readable media. For example, FIG. 12 shows a computer system (1200) suitable for implementing certain embodiments of the disclosed subject matter.

The computer software can be coded using any suitable machine code or computer language, that may be subject to assembly, compilation, linking, or like mechanisms to create code comprising instructions that can be executed directly, or through interpretation, micro-code execution, and the like, by one or more computer central processing units (CPUs), Graphics Processing Units (GPUs), and the like.

The instructions can be executed on various types of computers or components thereof, including, for example, personal computers, tablet computers, servers, smartphones, gaming devices, internet of things devices, and the like.

The components shown in FIG. 12 for computer system (1200) are exemplary in nature and are not intended to suggest any limitation as to the scope of use or functionality of the computer software implementing embodiments of the present disclosure. Neither should the configuration of components be interpreted as having any dependency or requirement relating to any one or combination of components illustrated in the exemplary embodiment of a computer system (1200).

Computer system (1200) may include certain human interface input devices. Such a human interface input device may be responsive to input by one or more human users through, for example, tactile input (such as: keystrokes, swipes, data glove movements), audio input (such as: voice, clapping), visual input (such as: gestures), olfactory input (not depicted). The human interface devices can also be used to capture certain media not necessarily directly related to conscious input by a human, such as audio (such as: speech, music, ambient sound), images (such as: scanned images, photographic images obtain from a still image camera), video (such as two-dimensional video, three-dimensional video including stereoscopic video).

Input human interface devices may include one or more of (only one of each depicted): keyboard (1201), mouse (1202), trackpad (1203), touch screen (1210), data-glove (not shown), joystick (1205), microphone (1206), scanner (1207), camera (1208).



Computer system (1200) may also include certain human interface output devices. Such human interface output devices may be stimulating the senses of one or more human users through, for example, tactile output, sound, light, and smell/taste. Such human interface output devices may include tactile output devices (for example tactile feedback by the touch-screen (1210), data-glove (not shown), or joystick (1205), but there can also be tactile feedback devices that do not serve as input devices), audio output devices (such as: speakers (1209), headphones (not depicted)), visual output devices (such as screens (1210) to include CRT screens, LCD screens, plasma screens, OLED screens, each with or without touch-screen input capability, each with or without tactile feedback capability—some of which may be capable to output two dimensional visual output or more than three dimensional output through means such as stereographic output; virtual-reality glasses (not depicted), holographic displays and smoke tanks (not depicted)), and printers (not depicted).

Computer system (1200) can also include human accessible storage devices and their associated media such as optical media including CD/DVD ROM/RW (1220) with CD/DVD or the like media (1221), thumb-drive (1222), removable hard drive or solid state drive (1223), legacy magnetic media such as tape and floppy disc (not depicted), specialized ROM/ASIC/PLD based devices such as security dongles (not depicted), and the like.

Those skilled in the art should also understand that term “computer readable media” as used in connection with the presently disclosed subject matter does not encompass transmission media, carrier waves, or other transitory signals.

Computer system (1200) can also include an interface to one or more communication networks. Networks can for example be wireless, wireline, optical. Networks can further be local, wide-area, metropolitan, vehicular and industrial, real-time, delay-tolerant, and so on. Examples of networks include local area networks such as Ethernet, wireless LANs, cellular networks to include GSM, 3G, 4G, 5G, LTE and the like, TV wireline or wireless wide area digital networks to include cable TV, satellite TV, and terrestrial broadcast TV, vehicular and industrial to include CANBus, and so forth. Certain networks commonly require external network interface adapters that attached to certain general purpose data ports or peripheral buses (1249) (such as, for example USB ports of the computer system (1200)); others are commonly integrated into the core of the computer system (1200) by attachment to a system bus as described below (for example Ethernet interface into a PC computer system or cellular network interface into a smartphone computer system). Using any of these networks, computer system (1200) can communicate with other entities. Such communication can be uni-directional, receive only (for example, broadcast TV), uni-directional send-only (for example CANbus to certain CANbus devices), or bi-directional, for example to other computer systems using local or wide area digital networks. Certain protocols and protocol stacks can be used on each of those networks and network interfaces as described above.

Aforementioned human interface devices, human-accessible storage devices, and network interfaces can be attached to a core (1240) of the computer system (1200).

The core (1240) can include one or more Central Processing Units (CPU) (1241), Graphics Processing Units (GPU) (1242), specialized programmable processing units in the form of Field Programmable Gate Areas (FPGA) (1243), hardware accelerators for certain tasks (1244), and so forth. These devices, along with Read-only memory

(ROM) (1245), Random-access memory (1246), internal mass storage such as internal non-user accessible hard drives, SSDs, and the like (1247), may be connected through a system bus (1248). In some computer systems, the system bus (1248) can be accessible in the form of one or more physical plugs to enable extensions by additional CPUs, GPU, and the like. The peripheral devices can be attached either directly to the core’s system bus (1248), or through a peripheral bus (1249). Architectures for a peripheral bus include PCI, USB, and the like.

CPUs (1241), GPUs (1242), FPGAs (1243), and accelerators (1244) can execute certain instructions that, in combination, can make up the aforementioned computer code. That computer code can be stored in ROM (1245) or RAM (1246). Transitional data can be also be stored in RAM (1246), whereas permanent data can be stored for example, in the internal mass storage (1247). Fast storage and retrieve to any of the memory devices can be enabled through the use of cache memory, that can be closely associated with one or more CPU (1241), GPU (1242), mass storage (1247), ROM (1245), RAM (1246), and the like.

The computer readable media can have computer code thereon for performing various computer-implemented operations. The media and computer code can be those specially designed and constructed for the purposes of the present disclosure, or they can be of the kind well known and available to those having skill in the computer software arts.

As an example and not by way of limitation, the computer system having architecture (1200), and specifically the core (1240) can provide functionality as a result of processor(s) (including CPUs, GPUs, FPGA, accelerators, and the like) executing software embodied in one or more tangible, computer-readable media. Such computer-readable media can be media associated with user-accessible mass storage as introduced above, as well as certain storage of the core (1240) that are of non-transitory nature, such as core-internal mass storage (1247) or ROM (1245). The software implementing various embodiments of the present disclosure can be stored in such devices and executed by core (1240). A computer-readable medium can include one or more memory devices or chips, according to particular needs. The software can cause the core (1240) and specifically the processors therein (including CPU, GPU, FPGA, and the like) to execute particular processes or particular parts of particular processes described herein, including defining data structures stored in RAM (1246) and modifying such data structures according to the processes defined by the software. In addition or as an alternative, the computer system can provide functionality as a result of logic hardware or otherwise embodied in a circuit (for example: accelerator (1244)), which can operate in place of or together with software to execute particular processes or particular parts of particular processes described herein. Reference to software can encompass logic, and vice versa, where appropriate. Reference to a computer-readable media can encompass a circuit (such as an integrated circuit (IC)) storing software for execution, a circuit embodying logic for execution, or both, where appropriate. The present disclosure encompasses any suitable combination of hardware and software.

#### APPENDIX A: ACRONYMS

JEM: joint exploration model  
 VVC: versatile video coding  
 BMS: benchmark set  
 MV: Motion Vector



HEVC: High Efficiency Video Coding  
 SEI: Supplementary Enhancement Information  
 VUI: Video Usability Information  
 GOPs: Groups of Pictures  
 TUs: Transform Units,  
 PUs: Prediction Units  
 CTUs: Coding Tree Units  
 CTBs: Coding Tree Blocks  
 PBs: Prediction Blocks  
 HRD: Hypothetical Reference Decoder  
 SNR: Signal Noise Ratio  
 CPUs: Central Processing Units  
 GPUs: Graphics Processing Units  
 CRT: Cathode Ray Tube  
 LCD: Liquid-Crystal Display  
 OLED: Organic Light-Emitting Diode  
 CD: Compact Disc  
 DVD: Digital Video Disc  
 ROM: Read-Only Memory  
 RAM: Random Access Memory  
 ASIC: Application-Specific Integrated Circuit  
 PLD: Programmable Logic Device  
 LAN: Local Area Network  
 GSM: Global System for Mobile communications  
 LTE: Long-Term Evolution  
 CANBus: Controller Area Network Bus  
 USB: Universal Serial Bus  
 PCI: Peripheral Component Interconnect  
 FPGA: Field Programmable Gate Areas  
 SSD: solid-state drive  
 IC: Integrated Circuit  
 CU: Coding Unit

While this disclosure has described several exemplary embodiments, there are alterations, permutations, and various substitute equivalents, which fall within the scope of the disclosure. It will thus be appreciated that those skilled in the art will be able to devise numerous systems and methods which, although not explicitly shown or described herein, embody the principles of the disclosure and are thus within the spirit and scope thereof.

What is claimed is:

1. A method for video decoding in a decoder, comprising:  
 decoding a first portion of video data to obtain first  
 decoded data corresponding to at least two pictures of  
 a plurality of pictures, the video data corresponding to  
 the plurality of pictures that is associated with respec-  
 tive Picture Order Count (POC) numbers indicating a  
 temporal order of the plurality of pictures and respec-  
 tively in N temporal layers;  
 allocating M memory spaces, M being a positive integer  
 less than or equal to N;  
 identifying one or more pictures of the at least two  
 pictures for decoding a second portion of the video data  
 corresponding to a current picture;  
 in a case that the one or more identified pictures includes  
 two or more identified pictures, selecting a collocated  
 reference picture from the two or more identified  
 pictures based on one of (i) the POC numbers of the  
 two or more identified pictures and the current picture,  
 and (ii) a selection index provided in the video data;  
 in a case that the one or more identified pictures includes  
 only one identified picture, selecting the only one  
 identified picture as the collocated reference picture;  
 decoding the second portion of the video data using the  
 collocated reference picture to obtain second decoded  
 data corresponding to the current picture;

storing the first decoded data corresponding to the one or  
 more identified pictures in one or more respective  
 memory spaces of the M allocated memory spaces; and  
 in a case that the one or more identified pictures includes  
 a particular picture in a same temporal layer as the  
 current picture, at least partially updating data stored in  
 one of the M memory spaces that corresponds to the  
 particular picture using the second decoded data cor-  
 responding to the current picture.

2. The method of claim 1, wherein, in the case that the one  
 or more identified pictures includes the two or more iden-  
 tified pictures, the selecting the collocated reference picture  
 from the two or more identified pictures comprises:

selecting one of the two or more identified pictures that  
 corresponds to a closest POC number difference with  
 respect to the current picture as the collocated reference  
 picture.

3. The method of claim 1, wherein, in the case that the one  
 or more identified pictures includes the two or more iden-  
 tified pictures, the selecting the collocated reference picture  
 from the two or more identified pictures is performed based  
 on the POC numbers of the two or more identified pictures  
 and the current picture, comprises:

in a case that two of the two or more identified pictures  
 correspond to a closest POC number difference with  
 respect to the current picture, performing one of  
 selecting one of the two of the two or more identified  
 pictures in a lowest temporal layer of the identified  
 pictures as the collocated reference picture, and  
 selecting one of the two of the two or more identified  
 pictures in a highest temporal layer of the identified  
 pictures as the collocated reference picture.

4. The method of claim 1, wherein  
 the M memory spaces are associated with M of the N  
 temporal layers, respectively.

5. The method of claim 4, further comprising:  
 obtaining the positive integer M from the video data.

6. The method of claim 4, further comprising:  
 determining the positive integer M based on a picture size  
 of the plurality of pictures.

7. The method of claim 4, wherein  
 the positive integer N is greater than two,  
 the positive integer M is less than or equal to the positive  
 integer N, and  
 the M allocated memory spaces are associated with M  
 lower temporal layers among the N temporal layers.

8. The method of claim 1, further comprising:  
 in a case that the one or more identified pictures includes  
 less than M pictures and the current picture is in a  
 lowest temporal layer of the N temporal layers, storing  
 the second decoded data corresponding to the current  
 picture in a vacant one of the M memory spaces.

9. The method of claim 8, further comprising:  
 in a case that the one or more identified pictures includes  
 M pictures, the one or more identified pictures and the  
 current picture are in different temporal layers, and the  
 one or more identified pictures include two pictures in  
 the lowest temporal layer, storing the second decoded  
 data corresponding to the current picture in one of the  
 M memory spaces that stores decoded data correspond-  
 ing to one of the two pictures in the lowest temporal  
 layer that has a smallest POC number.

10. The method of claim 1, wherein  
 the at least partially updating the data stored in the one of  
 the M memory spaces comprises storing the second



31

decoded data corresponding to the current picture in place of the decoded data corresponding to the particular picture.

**11.** An apparatus, comprising:

processing circuitry configured to:

decode a first portion of video data to obtain first decoded data corresponding to at least two pictures of a plurality of pictures, the video data corresponding to the plurality of pictures that is associated with respective Picture Order Count (POC) numbers indicating a temporal order of the plurality of pictures and respectively in N temporal layers;

allocate M memory spaces, M being a positive integer less than or equal to N;

identify one or more pictures of the at least two pictures for decoding a second portion of the video data corresponding to a current picture;

in a case that the one or more identified pictures includes two or more identified pictures, select a collocated reference picture from the two or more identified pictures based on one of (i) the POC numbers of the two or more identified pictures and the current picture, and (ii) a selection index provided in the video data;

in a case that the one or more identified pictures includes only one identified picture, select the only one identified picture as the collocated reference picture;

decode the second portion of the video data using the collocated reference picture to obtain second decoded data corresponding to the current picture;

store the first decoded data corresponding to the one or more identified pictures in one or more respective memory spaces of the M allocated memory spaces; and

in a case that the one or more identified pictures includes a particular picture in a same temporal layer as the current picture, at least partially update data stored in one of the M memory spaces that corresponds to the particular picture using the second decoded data corresponding to the current picture.

**12.** The apparatus of claim 11, wherein the processing circuitry is further configured to:

in the case that the one or more identified pictures includes the two or more identified pictures, select one of the two or more identified pictures that corresponds to a closest POC number difference with respect to the current picture as the collocated reference picture.

**13.** The apparatus of claim 11, wherein the processing circuitry is further configured to:

in a case that the one or more identified pictures include the two or more identified pictures, and two of the two or more identified pictures correspond to a closest POC number difference with respect to the current picture, perform one of

selecting one of the two of the two or more identified pictures in a lowest temporal layer of the identified pictures as the collocated reference picture, and

selecting one of the two of the two or more identified pictures in a highest temporal layer of the identified pictures as the collocated reference picture.

**14.** The apparatus of claim 11, wherein the M memory spaces are associated with M of the N temporal layers, respectively.

32

**15.** The apparatus of claim 14, wherein the positive integer N is greater than two, the positive integer M is less than or equal to the positive integer N, and

the M allocated memory spaces are associated with M lower temporal layers among the N temporal layers.

**16.** The apparatus of claim 11, wherein the processing circuitry is further configured to:

in a case that the one or more identified pictures includes less than M pictures and the current picture is in a lowest temporal layer of the N temporal layers, store the second decoded data corresponding to the current picture in a vacant one of the M memory spaces.

**17.** The apparatus of claim 16, wherein the processing circuitry is further configured to:

in a case that the one or more identified pictures includes M pictures, the one or more identified pictures and the current picture are in different temporal layers, and the one or more identified pictures include two pictures in the lowest temporal layer, store the second decoded data corresponding to the current picture in one of the M memory spaces that stores decoded data corresponding to one of the two pictures in the lowest temporal layer that has a smallest POC number.

**18.** A non-transitory computer-readable medium storing instructions which when executed by a computer for video decoding causes the computer to perform:

decoding a first portion of video data to obtain first decoded data corresponding to at least two pictures of a plurality of pictures, the video data corresponding to the plurality of pictures that is associated with respective Picture Order Count (POC) numbers indicating a temporal order of the plurality of pictures and respectively in N temporal layers;

allocating M memory spaces, M being a positive integer less than or equal to N;

identifying one or more pictures of the at least two pictures for decoding a second portion of the video data corresponding to a current picture;

in a case that the one or more identified pictures include two or more identified pictures, selecting a collocated reference picture from the two or more identified pictures based on one of (i) the POC numbers of the two or more identified pictures and the current picture, and (ii) a selection index provided in the video data;

in a case that the one or more identified pictures includes only one identified picture, selecting the only one identified picture as the collocated reference picture;

decoding the second portion of the video data using the collocated reference picture to obtain second decoded data corresponding to the current picture;

storing the first decoded data corresponding to the one or more identified pictures in one or more respective memory spaces of the M allocated memory spaces; and

in a case that the one or more identified pictures includes a particular picture in a same temporal layer as the current picture, at least partially updating data stored in one of the M memory spaces that corresponds to the particular picture using the second decoded data corresponding to the current picture.

**19.** The non-transitory computer-readable medium of claim 18, wherein

the M memory spaces are associated with M of the N temporal layers, respectively.

\* \* \* \* \*